FERMILAB-PUB-25-0031-CSAID

SIDDA: SInkhorn Dynamic Domain Adaptation for Image Classification with Equivariant Neural Networks

Sneh Pandya

Department of Physics, Northeastern University, Boston, Massachusetts 02115
NSF AI Institute for Artificial Intelligence & Fundamental Interactions (IAIFI), Cambridge, MA 02139
Fermi National Accelerator Laboratory, Batavia, IL 60510

Purvik Patel

Khoury College of Computer Science, Northeastern University, Boston, Massachusetts 02115

Brian D. Nord

Fermi National Accelerator Laboratory, Batavia, IL 60510
Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637
Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637

Mike Walmsley

Dunlap Institute for Astronomy & Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, University of Manchester, Manchester, M13 9PL, UK

Aleksandra Ćiprijanović

Fermi National Accelerator Laboratory, Batavia, IL 60510
Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637
NSF-Simons AI Institute for the Sky (SkAI), 172 E. Chestnut St., Chicago, IL 60611, USA

Abstract. Modern neural networks (NNs) often do not generalize well in the presence of a “covariate shift”; that is, in situations where the training and test data distributions differ, but the conditional distribution of classification labels given the data remains unchanged. In such cases, NN generalization can be reduced to a problem of learning more robust, domain-invariant features. Domain adaptation (DA) methods include a broad range of techniques aimed at achieving this; however, these methods have struggled with the need for extensive hyperparameter tuning, which then incurs significant computational costs. In this work, we introduce SIDDA, an out-of-the-box DA training algorithm built upon the Sinkhorn divergence, that can achieve effective domain alignment with minimal hyperparameter tuning and computational overhead. We demonstrate the efficacy of our method on multiple simulated and real datasets of varying complexity, including simple shapes, handwritten digits, and real astronomical observations. These datasets exhibit covariate shifts due to noise, blurring, and differences between telescopes. SIDDA is compatible with a variety of NN architectures, and it works particularly well in improving classification accuracy and model calibration when paired with symmetry-aware equivariant neural networks (ENNs). We find that SIDDA consistently enhances the generalization capabilities of NNs, achieving up to a $\approx 40\%$ improvement in classification accuracy on unlabeled target data, while also providing a more modest performance gain of $\lesssim 1\%$ on labeled source data. We also study the efficacy of DA on ENNs with respect to the varying group orders of the dihedral group D_N, and find that the model performance improves as the degree of equivariance increases. Finally, we find that SIDDA enhances model calibration on both source and target data, with the most significant gains in the unlabeled target domain—achieving over an order of magnitude improvement in the expected calibration error and Brier score. SIDDA's versatility across various NN models and datasets, combined with its automated approach to domain alignment, has the potential to significantly advance multi-dataset studies by enabling the development of highly generalizable models.

1. Introduction

Deep neural networks (NNs) excel at extracting complex features from data, making them a powerful tool for a wide range of tasks, including classification, regression, and anomaly detection. Unfortunately, some extracted features can be very dataset-specific, which makes it challenging for NN models to generalize to data that differs from the training data, even when the differences are subtle. For instance, a significant drop in performance occurs when the input distribution changes between the training and test datasets, despite the conditional distribution of the labels given the inputs remaining the same — a scenario commonly referred to as a "covariate shift" [Farahani et al., 2020, Liu et al., 2022].

Generalization allows models to perform well across diverse data domains, ranging from subtle variations in input distributions to entirely different datasets or environments. Differences between training and testing data can be due to data origin or quality [Dodge and Karam, 2017, Gide et al., 2016, Dodge and Karam, 2016, Ford et al., 2019], or even single-pixel level differences, which can cause the NN to give inaccurate predictions [Su et al., 2019]. Generalization capabilities, in turn, aid the efficiency and applicability of NNs in both science and industry, as they would otherwise need to be continually retrained on new data. For example, in astronomy, a generalized model trained on data from one telescope should accurately predict properties of data from another telescope that has different noise characteristics or resolution, significantly accelerating the process of identifying or characterizing celestial objects across surveys.

Domain adaptation (DA) is a group of methods that aim to improve the generalization capabilities of NNs by enabling the NN to learn features in the data that persist across domains [Wang and Deng, 2018, Csurka, 2017, Wilson and Cook, 2020]. It is often applied to problems where one has access to labeled data from a "source" domain, but would also like the model to perform well on unlabeled "target" domain data. A large group of distance-based DA methods tackles the covariate shift problem by minimizing some distance metric between internal NN latent representations (distributions) of the source and target data. This, in turn, forces the NN to extract mainly domain-invariant features, which makes both latent data distributions well-aligned.

Some well-known distance-based DA methods use maximum mean discrepancy (MMD) [Gretton et al., 2008], correlation alignment (CORAL) [Sun et al., 2016], contrastive domain discrepancy (CDD) [Kang et al., 2019], the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951], or the Wasserstein distance [Panaretos and Zemel, 2018], which is derived from the optimal transport (OT) [Courty et al., 2014] theory for measuring distance between probability distributions. The theory is aimed at solving the OT problem, which includes determining the minimal cost of transporting a probability mass from one distribution to another, where the cost is defined by a chosen metric that measures the effort required to move it. OT thus provides a principled way to quantify the dissimilarity between distributions, capturing both the geometry of the space and the magnitude of the differences.

In the sciences, NNs have made significant progress — from mapping the structure of biological proteins [Jumper et al., 2021] to the cosmos [Jeffrey et al., 2024]. In astronomy, astrophysics, and cosmology, the success is mainly due to the emergence of large datasets and high-fidelity simulations. Stage IV projects, such as the Vera Rubin Observatory Legacy Survey of Space and Time (LSST) [Ivezić et al., 2019], the Nancy Grace Roman Telescope [Akeson et al., 2019], and the Euclid mission [Scaramella et al., 2022], will yield an unprecedented amount of data that analysis pipelines must analyze efficiently. Simultaneously, there are many simulations of the Universe from subparsec to gigaparsec scales with magneto-hydrodynamic simulation suites, such as IllustrisTNG [Nelson et al., 2021] and CAMELS [Villaescusa-Navarro et al., 2021], that can be used to prepare pipelines for the analysis of real data. Since simulation-trained models often exhibit a substantial drop in performance when applied to real data, there has been extensive work in implementing DA for astronomical applications [Vilalta et al., 2019, Ćiprijanović et al., 2022, 2023, Roncoli et al., 2023, Swierc et al., 2023, Gilda et al., 2024, Agarwal et al., 2024, Parul et al., 2024], where atmospheric distortion, telescope noise, PSF blurring, and data processing errors commonly affect data quality.

Most image-based problems in the sciences are addressed using various types of convolutional neural networks (CNNs). At the same time, equivariant neural networks (ENNs) are gaining popularity due to their ability to explicitly encode symmetry information present in the data, which is often explicitly known — e.g., SL$(2, \mathbb{C})$ in particle physics [Bogatskiy et al., 2020] and SE(3) for rigid body motion [Fuchs et al., 2020]. ENNs have also been shown to achieve state-of-the-art performance on many tasks [Esteves et al., 2018, Satorras et al., 2021, Deng et al., 2021, Kalogeropoulos et al., 2024] and to possess inherent robustness to symmetric transformations and noise perturbations due to their restricted feature learning [Pandya et al., 2023, Bulusu et al., 2022, Du et al., 2022]. This robustness has been observed to increase with the group order N for the cyclic group C_N and dihedral group D_N, which are subgroups of SO(2) and O(2), respectively. Still, in the presence of covariate shifts, even ENNs can exhibit a drop in performance [Pandya et al., 2023].

In this work, we focus on the Sinkhorn divergence [Altschuler et al., 2018], a symmetrized variant of regularized OT distances. We introduce SInkhorn Dynamic Domain Adaptation (SIDDA), a more automated training algorithm for DA that minimizes the need for hyperparameter tuning. To achieve this, we leverage active scaling of (1) the entropic regularization of the OT plan, and (2) the weighting of classification and DA loss terms, during training. To demonstrate the efficacy and broad scope of applications of our proposed DA method, in this work, we use several datasets of varying complexity: simple simulated datasets, well-known benchmark datasets used in the computer science community, and real observational galaxy datasets. We also study the robustness of ENNs and the improved efficacy of SIDDA when used in conjunction with ENNs compared to typical CNNs.

The paper is organized as follows. In Section 2, we describe existing DA methods and their shortcomings, motivating the use of OT-based distances to facilitate DA. We describe the core methodology of SIDDA, and motivate the use of ENNs as inherently robust architectures. In Section 3, we describe the construction of all simulated and real datasets we use in our study. In Section 4, we describe the network architectures used, training procedures, metrics for calibration, and metrics for interpreting NN latent spaces. In Section 5, we summarize our results and conclude in Section 6.

2. Methods

The major components of our work are DA and equivariance, which we combine to create a more efficient and robust NN classifier. Within DA, we implement the Sinkhorn divergence, a symmetrized and regularized variant of OT distances that offers considerable improvement in DA over traditional methods. We construct a training program that constantly adjusts the loss landscape and regularization strength of the Sinkhorn plan, offering optimal domain alignment with minimal hyperparameter tuning.

2.1. Domain Adaptation

DA comprises a set of techniques aimed at aligning the latent distributions of NNs in the presence of covariate shifts in data. Typically, DA operates in settings where one can access labeled source images $x \in \mathbf{X_s} \subseteq \mathbb{R}^{m\times m}$, and unlabeled target images $x^* \in \mathbf{X_t} \subseteq \mathbb{R}^{m\times m}$ from $\mathbf{X}_s$ and $\mathbf{X}_t$ source and target data domains, where m denotes the number of pixels in each dimension (height and width) of the image.

Consider the latent vectors $z \in \mathbf{Z}_s \subseteq \mathbb{R}^l$ and $z^* \in \mathbf{Z}_t \subseteq \mathbb{R}^l$, where $\mathbf{Z}_s$ and $\mathbf{Z}_t$ denote the latent spaces of the source and target domains, respectively, and l represents the dimension of the latent vectors (i.e., the width of the corresponding neural network layer). Latent distributions refer to the probability distributions over these latent vectors, and during training, DA minimizes a statistical distance measure between them. DA is incorporated through an additional loss term, $\mathcal{L}_{\text{DA}}$, alongside the standard task loss (e.g., cross-entropy for classification, $\mathcal{L}_{\text{CE}}$), to promote alignment between the two latent distributions. The total loss function is then:

$$\mathcal{L} \propto \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DA}} \ . \tag{1}$$

In practice, a delicate balance between the two terms must be achieved to ensure proper alignment.

There are numerous DA methods, each with its own strengths and limitations. One commonly used approach is MMD [Gretton et al., 2008, 2012], where the distance between the means of the latent embeddings from the source and target domains serves as the DA loss function. In DA, comparisons are often made between distributions that are not explicitly known but can be sampled. MMD can be combined with kernel methods, which map probability distributions into a high-dimensional reproducing kernel Hilbert space (RKHS) [Gretton et al., 2012], providing a more flexible method for comparing distributions. This approach allows for analyzing distributions through well-defined operations in the RKHS, even when the original distributions are not well-defined. The MMD between two probability distributions μ and ν — representing distributions over latent vectors z and z^*, respectively — is

$$\text{MMD}(\mu,\nu) = \Bigg(\mathbb{E}_{z,z'\sim\mu}\left[k(z,z')\right] + \mathbb{E}_{z^*,z^{*'}\sim\nu}\left[k(z^*,z^{*'})\right] - 2\mathbb{E}_{z\sim\mu,z^*\sim\nu}\left[k(z,z^*)\right]\Bigg)^{1/2} , \tag{2}$$

where k represents the kernel function, and z, z', and z^*, $z^{*'}$ are individual samples from latent distributions μ and ν, respectively.

Despite its utility, MMD has several theoretical and implementation-related shortcomings. First, its efficacy is highly sensitive to the choice of k. In typical applications, the Gaussian kernel $k(z,z^*) = \exp\left(-\frac{||z-z^*||^2}{2\epsilon^2}\right)$ is used with kernel bandwidth ϵ. Other kernel options include the linear kernel $k(z,z^*) = z^T z^*$, the Laplacian kernel $k(z,z^*) = \exp\left(-\frac{||z-z^*||}{2\epsilon}\right)$, and others. Most kernels generally belong to a one-parameter family (e.g., ϵ for Gaussian and Laplacian kernels) and must be carefully tuned, or complex linear combinations of kernels with many different parameter values must be used. The specific choice of kernel depends heavily on the nature of the problem. That is, MMD can exhibit bias with small sample sizes and often struggles with domain alignment when dealing with high-dimensional distributions [Muandet et al., 2017, Reddi et al., 2014].

2.2. Optimal Transport and The Sinkhorn Divergence

OT distances and their symmetrized variants, such as Sinkhorn divergences, offer an alternative to MMD. Traditionally, computing OT is prohibitively expensive [Peyré and Cuturi, 2020]. Entropic regularization, OT_σ, [Dessein et al., 2018] provides a more efficient method for estimating OT distances. The regularized OT is defined as

$$\mathrm{OT}_\sigma(\mu,\nu) = \min_{\gamma\in U(\mu,\nu)} \left(\sum_{i,j} \gamma_{ij} d(z_i, z_j^*)^p + \sigma H(\gamma) \right) , \tag{3}$$

where $d(z_i, z_j^*)^p$ is the distance between source feature z_i and target feature z_j^*. When $p = 1$, this distance becomes the Earth Mover's distance [Rubner et al., 1998], and when $p = 2$, it becomes the quadratic Wasserstein distance. The transport plan $\gamma \in U(\mu,\nu)$ is a joint probability distribution between μ and ν, where the set of admissible transport plans $U(\mu,\nu)$ is defined by the marginal constraints:

$$\sum_j \gamma_{ij} = \mu_i, \quad \sum_i \gamma_{ij} = \nu_j. \tag{4}$$

The entropy $H(\gamma) = -\sum_{i,j} \gamma_{ij} \log \gamma_{ij}$ regularizes the transport plan γ, and σ controls the regularization strength. One limitation of OT_σ is that $\mathrm{OT}_\sigma(\mu,\mu) \neq 0$, implying a non-zero cost even when transporting a distribution to itself, leading to bias in the measure.

To correct this bias, the Sinkhorn divergence $S_\sigma(\mu,\nu)$, defined as

$$S_\sigma(\mu,\nu) = \mathrm{OT}_\sigma(\mu,\nu) - \frac{1}{2}\mathrm{OT}_\sigma(\mu,\mu) - \frac{1}{2}\mathrm{OT}_\sigma(\nu,\nu), \tag{5}$$

can compensate for the bias in OT_σ [Feydy et al., 2018]. As $\sigma \to 0$, $S_\sigma(\mu,\nu)$ converges to the (biased) optimal transport OT_0, and as $\sigma \to \infty$, it interpolates towards MMD loss [Feydy et al., 2018]. For small values of σ, an unbiased transport plan that still enjoys the benefits of OT-based distances can be constructed.

2.3. Dynamic Sinkhorn Divergences for Domain Adaptation

For this work, $\mathcal{L}_{\mathrm{DA}}$ in Equation 1 is specifically the Sinkhorn divergence $S_\sigma(\mu,\nu)$. However, a careful balance between $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{DA}}$ must be achieved to optimize the classification task while simultaneously maximizing domain alignment.

Finding the best weights for each of the loss terms can be very challenging and time-consuming. Furthermore, a single choice of weights might not be the best choice throughout the whole training procedure. To manage this balance, we employ dynamic weighting of the losses by introducing two trainable parameters, η_1 and η_2, which dynamically adjust the contributions of the loss terms for each task. These parameters ensure that no single loss term dominates the optimization process, allowing the loss landscape to be optimally adjusted for both tasks. Drawing inspiration from Kendall et al. [2018], we use the following for the total loss function:

$$\mathcal{L} = \frac{1}{2\eta_1^2}\mathcal{L}_{\mathrm{CE}} + \frac{1}{2\eta_2^2}\mathcal{L}_{\mathrm{DA}} + \log(|\eta_1\eta_2|) , \tag{6}$$

where η_1 and η_2 are trainable scalars, and their values are jointly learned with the model weights during training. The inclusion of the term $\log(|\eta_1\eta_2|)$ acts as a regularization to prevent η_1 and η_2 from collapsing to unstable values, such as zero. As $\eta_i \to 0$, the corresponding loss term is more heavily weighted. To ensure that no single component dominates, we impose the additional constraint $\eta_2 \geq \frac{\eta_1}{4}$. In general, the DA term must not dominate over the classification loss, which the above inequality enforces. For our implementation, we found that this threshold worked best, but such a cutoff may not always be optimal.

In Kendall et al. [2018], the two weight terms η_1 and η_2 were introduced for the dynamic weighting of the losses. These terms explicitly minimize the regression uncertainty associated with each loss term, as their model outputs a Gaussian distribution with variance η_i^2 for each task. In the case of classification, their weight terms become $1/\eta_i^2$. Since uncertainties are not one of the network outputs in our case, the exact written form of loss weights is not important and the extra factor of two can very well be absorbed into the trainable weight parameter.

The level of regularization σ in $S_\sigma(\mu,\nu)$ is another critical hyperparameter [Feydy et al., 2018]. When σ is too small, the transport plan's entropic regularization diminishes, and the OT plan's bias makes it susceptible to overfitting to specific sample locations, hindering the overlap between μ and μ latent distributions. Conversely, if σ is too large, the regularization interpolates toward MMD, removing the unique benefits of using $S_\sigma(\mu,\nu)$. To address this, we adopt a unique, dynamic regularization per epoch of training ℓ, σ_ℓ, where the transport plan is continually updated. We compute σ_ℓ iteratively as:

$$\sigma_\ell = \max\left(0.05 \cdot \max_{i,j} \|z_i - z_j^*\|_2,\ 0.01\right) . \tag{7}$$

In this formulation, σ_ℓ is dynamically adjusted based on the maximum pairwise distance $D_{ij} = ||z_i - z_j^*||_2$

between the source and target latent distributions. We additionally set scaling based on the appropriate measures on the unit square or cube, which justifies the prefactor of 0.05 in Equation 7 [Feydy et al., 2018]. This is further stabilized through the layer normalization of the latent vectors prior to computing D_{ij}. This stabilization discourages outliers from disproportionally affecting the computation of σ_ℓ. We also enforce a lower bound of $\sigma_\ell \geq 0.01$ as mitigation against the potential overfitting once the latent distributions have become sufficiently aligned.

Batches of size n of latent vectors from source and target data, denoted z_n and z_n^*, are retrieved through a forward pass of a single combined batch of the source and target data, $\mathbf{X} = [x_n, x_n^*]$, through the NN. The NN outputs a combined batch of latent vectors, denoted as $\mathbf{Z}$, which is subsequently separated into two subsets: one corresponding to the source domain and the other to the target domain. This is particularly important for NNs which utilize batch normalization [Ioffe and Szegedy, 2015]. If z_n and z_n^* were passed separately, the batch statistics would be computed independently, leading to inconsistent normalization as the z_n batch statistics will not incorporate z_n^* and vice versa.

We name this combined approach, which is dynamically adjusting a balance between cross-entropy and DA loss terms, and likewise dynamically adjusting the regularization of the Sinkhorn plan that facilitates the DA, SIDDA. SIDDA not only facilitates effective alignment between source and target domains, as we will demonstrate, but also reduces the need for extensive hyperparameter tuning — a common challenge in DA implementations [Saito et al., 2021]. As a result, this method provides an more automatic and reliable DA implementation with minimal computational overhead, leveraging existing resources that allow efficient computation of Sinkhorn divergences [Feydy et al., 2020]. We implement this technique using the `geomloss` library [Feydy et al., 2019], which provides GPU implementations for Sinkhorn divergences and compatibility with `PyTorch` [Paszke et al., 2019]. A pipeline illustrating the dynamic DA approach during training is given in Figure 2.3.

2.4. The Jensen-Shannon Distance

Recent advancements in DA theory have introduced the Jensen-Shannon (JS) divergence [Lin, 1991] as a fundamental tool for understanding the inherent limitations of DA [Shui et al., 2022]. The JS divergence is a symmetrized statistical distance metric. For two latent distributions μ and ν, the JS divergence D_{JS} is defined as

$$D_{\mathrm{JS}}(\mu\|\nu) = \frac{1}{2}D_{\mathrm{KL}}(\mu\|\tau) + \frac{1}{2}D_{\mathrm{KL}}(\nu\|\tau) \; , \tag{8}$$

where D_{KL} denotes the KL divergence [Kullback and Leibler, 1951], defined as

$$D_{\mathrm{KL}}(\mu \mid\mid \nu) = \int_{-\infty}^{\infty} p(z) \log\left(\frac{p(z)}{q(z)}\right) dz \; . \tag{9}$$

Here, p and q denote probability densities of the latent features z in the source and target distributions μ and ν. $\tau = \frac{1}{2}(\mu+\nu)$ represents the mixture distribution of μ and ν. The JS divergence offers two key advantages over D_{KL}: it is symmetric (since, in general, $D_{\mathrm{KL}}(\mu\|\nu) \neq D_{\mathrm{KL}}(\nu\|\mu)$), and it is always finite. Additionally, the square root $\sqrt{D_{\mathrm{JS}}}$ defines a metric known as the Jensen-Shannon distance.

For DA applications, there exists a lower bound on the target domain loss [Shui et al., 2022]:

$$\mathcal{L}_t(z^*) \geq \mathcal{L}_s(z) - \sqrt{D_{\mathrm{JS}}(\mu\|\nu)} \; , \tag{10}$$

where $\mathcal{L}_s$ is the source domain loss, $\mathcal{L}_t$ is the target domain loss, and $\sqrt{D_{\mathrm{JS}}(\mu\|\nu)}$ is the JS distance between the source and target latent distributions μ and ν, respectively. This bound emphasizes that perfect alignment between source and target domains is fundamentally constrained by two components: $\mathcal{L}_s$ and the JS distance between the source and target distributions. A smaller JS distance implies that the feature distributions of the source and target domains are closely aligned, which lowers the bound on $\mathcal{L}_t$ and enables better transferability of the learned model. Conversely, a larger JS distance indicates a greater discrepancy, limiting the potential for minimizing target domain loss through adaptation alone.

The similarity between the source and the target latent distributions μ and ν is inherently influenced by the feature extraction capabilities of the neural network. DA methods aim to align features in the latent distribution; however, these features are ultimately limited by the network architecture. As a toy example, consider the case of image classification, where the architecture is a multi-layer perceptron (MLP). Many image classification tasks exhibit translation invariance, inherent in CNNs, but not in MLPs. DA on this task with CNNs will likely be more successful than with MLPs, as the translation invariance of the CNN further restricts the allowable features, and thus, the cost of alignment will be smaller. In particular, we define "robust" features as those that respect the underlying data symmetries. More specifically, they yield similar classification probabilities under isometries that preserve the symmetries of the images. If these symmetries persist in both the source and target domain, which is typically true except for extreme symmetry-breaking perturbations, then the cost of aligning robust features will be less than features learned from symmetry-agnostic architectures. Consequently, it is reasonable to expect that ENNs

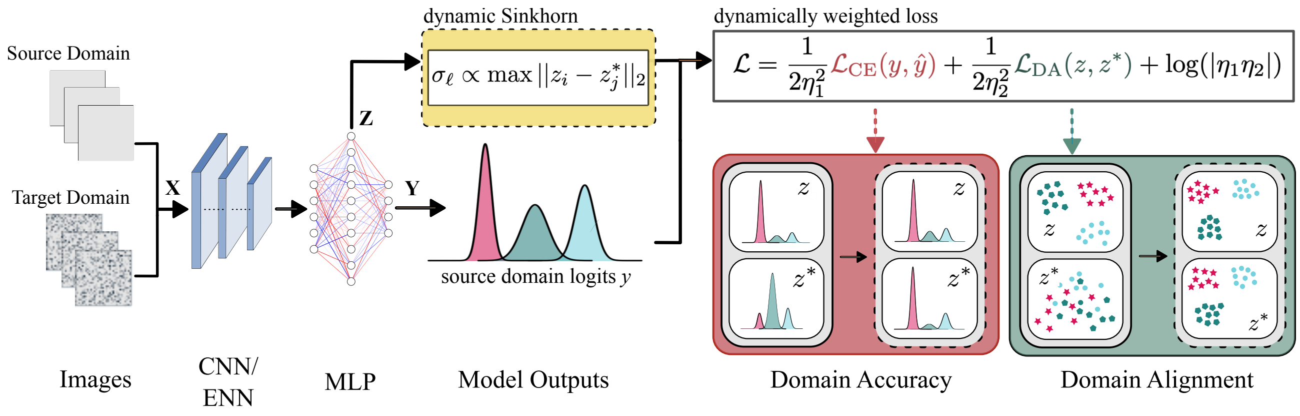
dynamic Sinkhorn
$\sigma_\ell \propto \max||z_i - z_j^*||_2$
dynamically weighted loss
$\mathcal{L} = \frac{1}{2\eta_1^2}\mathcal{L}_{\rm CE}(y, \hat{y}) + \frac{1}{2\eta_2^2}\mathcal{L}_{\rm DA}(z, z^*) + \log(|\eta_1\eta_2|)$
Source Domain
Target Domain
X
Z
Y
source domain logits y
z
z*
Images
CNN/ ENN
MLP
Model Outputs
Domain Accuracy
Domain Alignment

Figure 1. SIDDA pipeline. The source and target domain batches of size n, x_n and x_n^*, are first concatenated into a single batch $\mathbf{X}$ before being passed into the model. After passing through the convolutional layers, the neural network produces a combined batch of latent vectors, $\mathbf{Z}$, extracted from the final linear layer. This layer is positioned just before the output layer, which generates the class probabilities, $\mathbf{Y}$. Both $\mathbf{Z}$ and $\mathbf{Y}$ are split into separate batches for the source and target domains, resulting in z_n (source) and z_n^* (target) from $\mathbf{Z}$, and y_n (source) and y_n^* (target) from $\mathbf{Y}$, respectively. Only source y_n is used in training, as there are typically no target domain labels. Both z_n and z_n^* are used to compute σ_ℓ, a parameter that iteratively updates the regularization of the Sinkhorn plan in $\mathcal{L}_{\rm DA}$. This process aligns the latent distributions of the source and target domains. This loss contribution is appropriately weighted with the classification loss, $\mathcal{L}_{\rm CE}$, using a dynamic weighting of the tasks. The result of training using SIDDA is improved classification accuracy in both domains due to the aligned latent distributions, which can be visualized using non-linear clustering algorithms on the NN latent distributions.

Algorithm 1 SIDDA optimization step

```
Require: $x$: Source domain inputs, $\hat{y}$: Source domain labels, $x^*$: Target domain inputs, $n$: batch size in each of
    the domains, $\eta_1$, $\eta_2$: Dynamic weighting parameters
while not converged do
    $\mathbf{X} \leftarrow [x_n,\ x_n^*]$                                        // Concatenate inputs
    $\mathbf{Y},\ \mathbf{Z} \leftarrow$ model($\mathbf{X}$)                                // Compute logits and latent features
    $z_n,\ z_n^* \leftarrow \mathbf{Z}$                                          // Split features into source and target
    $y_n,\ y_n^* \leftarrow \mathbf{Y}$                                          // Split predictions into source and target
    $D_{ij} \leftarrow \|z_i - z_j^*\|_2,\ \forall\, i, j \in \{1, \ldots, n\}$                 // Compute pairwise distances
    $\sigma_\ell \leftarrow \max(0.05 \times \max_{i,j} D_{ij},\ 0.01)$                  // Compute dynamic blur parameter
    $\mathcal{L}_{\rm DA} \leftarrow$ Sinkhorn($z_n,\ z_n^*,\ \sigma_\ell$)                       // Compute Sinkhorn loss
    $\mathcal{L}_{\rm CE} \leftarrow$ CrossEntropy($y_n,\ \hat{y}_n$)                         // Compute classification loss
    $\mathcal{L} \leftarrow \frac{1}{2\eta_1^2}\mathcal{L}_{\rm CE} + \frac{1}{2\eta_2^2}\mathcal{L}_{\rm DA} + \log(|\eta_1\eta_2|)$          // Compute total loss
    loss.backward()                                           // Compute gradients
    clip_grad_norm_(model.parameters(), 10.0)                 // Gradient clipping
    $\eta_1 \leftarrow \max(\eta_1,\ 1e\text{-}3)$                                 // Bound $\eta_1$
    $\eta_2 \leftarrow \max(\eta_2,\ 0.25 \times \eta_1)$                          // Bound $\eta_2$
    optimizer.step()                                          // Update model parameters
    Evaluate convergence criteria                             // Check for early stopping
    if convergence criteria met then
        break                                                   // Early-stopping criterion
    end if
end while
```

endowed with appropriate higher-order symmetries will exhibit greater robustness and achieve more precise DA alignment than CNNs, because the latent distributions of ENNs are inherently more constrained. Furthermore, when the assumption of underlying symmetries holds in both the source and target domains, as is typical in many DA applications, this advantage of ENNs becomes even more pronounced.

3. Data

We evaluate the performance of our method on three simulated datasets and one real astronomical dataset: (1) a single-channel dataset of shapes consisting of lines, circles, and rectangles; (2) a single-channel dataset resembling astronomical objects, including stars, spirals, and elliptical galaxies; (3) the multichannel MNIST-M dataset [Ganin et al., 2016]; and (4) the Galaxy Zoo (GZ) Evo dataset of observed galaxies [Walmsley et al., 2024]. The shapes and astronomical objects datasets are constructed using `DeepBench` [Voetberg et al., 2023]. All datasets used in our experiments can be found on Zenodo.

3.1. Covariate Shifts

We use images from three simulated datasets, shown in Figure 2, to study the performance on induced covariate shifts between the source and target domains. For all of our simulated datasets, we introduce fixed levels of Poisson noise in the target domain. Additionally, for MNIST-M, we also study the effects of PSF blurring in the target domain. By studying these two distinct covariate shifts, we evaluate the robustness of our method on covariate shifts relevant to data in realistic settings, particularly in the context of astrophysics and cosmology.

This is implemented for an image I with grid values (ζ, ξ) and channels c as

$$I_{\text{Poisson}}(\zeta, \xi, c; S) = I(\zeta, \xi, c) + P\left(\frac{\langle I \rangle}{S} - \langle I \rangle\right) \,, \quad (11)$$

where S is the signal-to-noise ratio, and P denotes the Poisson distribution with rate parameter $\lambda = \frac{\langle I \rangle}{S} - \langle I \rangle$. We incur PSF noise in each image channel by convolving the images with a Gaussian kernel G of kernel width ϵ:

$$I_{\text{PSF}}(\zeta, \xi) = (I * G)(\zeta, \xi), \quad (12)$$

where

$$G(\zeta, \xi) = \frac{1}{2\pi\epsilon^2} \exp\left(-\frac{\zeta^2 + \xi^2}{2\epsilon^2}\right) \,. \quad (13)$$

3.2. Simulated Images

For the shapes dataset, we use `DeepBench` [Voetberg et al., 2023], an open-source library for generating simulated datasets, and randomly construct rectangles, lines, and circles with varying radii (for circles), heights and widths (for rectangles), and lengths (for lines). The object positions, orientations, and thickness are also randomly assigned to introduce variance in the dataset. Poisson noise in the images is normalized with respect to the original image signal. We set a signal-to-noise ratio $S = 0.05$. Example images from the dataset can be seen in the left two panels of Figure 2.

We also use `DeepBench` for simulating astronomical objects. We generate astronomical objects resembling spiral galaxies, elliptical galaxies, and stars. For spiral galaxies, we randomly assign the centroid, winding number, and pitch to ensure morphological variation. The pitch is the angle indicating how tightly the arms are wound, while the winding number measures the total number of arm rotations from the center to the galaxy's edge. For elliptical galaxies, we vary the amplitude, radius, ellipticity, Sérsic index [Sersic, 1958], and rotation, as well as the centroid location. The amplitude sets the brightness level, ellipticity describes the degree of deviation from a circle, the Sérsic index controls light concentration (higher values indicate more central concentration), and the rotation defines the orientation angle of the galaxy's major axis. Lastly, we apply similar variations to generate stars, with the number of stars in each image uniformly distributed in the range $[0, 10]$. We use a fixed Poisson noise level of $S = 0.2$ to generate noisy target domain images. This level of noise was chosen as it allows for a sufficient decrease in target domain performance for models without DA. Example images are shown in the middle two panels of Figure 2. Both of these datasets contain 12,000 training images (with 20% being used for validation) and 3,000 test images in each domain. The images are square with 100 pixels on each side, and they are single-channel: each sample image has dimensions $100 \times 100 \times 1$.

MNIST-M [Ganin et al., 2016] is a dataset that combines the handwritten digits of MNIST [Deng, 2012] with randomly extracted color photos from BSDS500 [Arbelaez et al., 2011] as background images. The original dataset contains 59,001 training images and 90,001 test images, out of which we use a balanced subset of 15,000 training (with 20% set aside for validation) and 5,000 testing images in each domain. Since this is a three-channel dataset, images have a dimension of $32 \times 32 \times 3$. We then create two types of target domain covariate shifts: 1) we set a signal-to-noise ratio of $S = 0.05$ for Poisson noise, and 2) a kernel width of $\epsilon = 2$ for PSF blurring. Example images are shown in the right two panels of Figure 2.

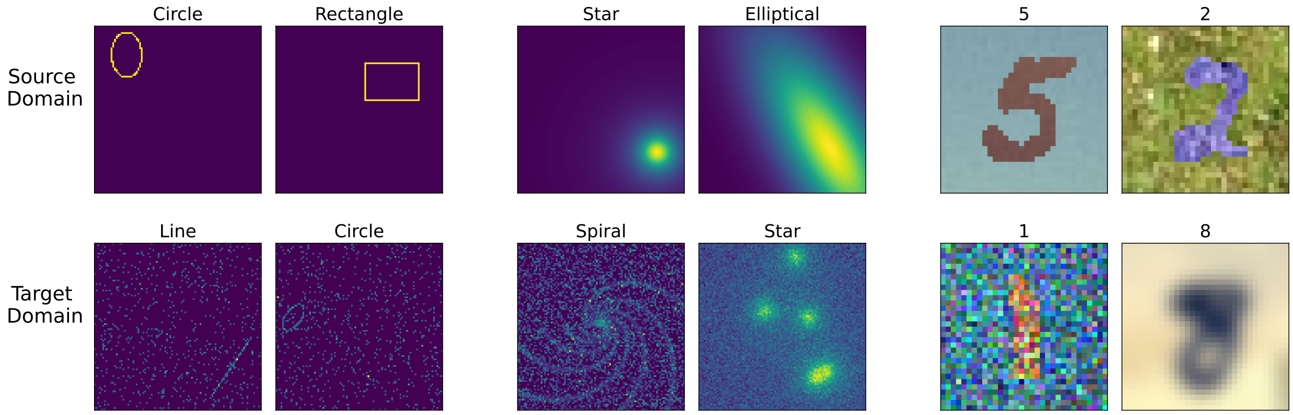

Figure 2. Example images for simulated datasets in the source domain (top row) and the target domain (bottom row) with corresponding labels. **Left Panels:** Shapes dataset, featuring lines, rectangles, and circles, simulated with `DeepBench`. This dataset includes variations in object positions and orientations, with Poisson noise added and normalized relative to the image signal in the target domain. **Middle Panels:** Astronomical objects dataset, generated using `DeepBench`. Parameters for spiral and elliptical galaxies were randomly sampled to determine morphology and position, while stars were generated similarly, with the number of stars as an additional parameter. Target domain images include additional Poisson noise. **Right Panels:** MNIST-M dataset with simulated Poisson noise (bottom left) and PSF blurring (bottom right) in the target domain.

Both the images and the induced covariate shifts are simulated and do not capture all the complexities of real-world noise. Nevertheless, these datasets provide valuable benchmarks for challenges commonly encountered in astronomical and cosmological contexts, where DA methods can substantially enhance the robustness of neural network-based image classification pipelines.

3.3. Real-Sky Galaxy Image Dataset

We use the GZ Evo dataset [Walmsley et al., 2024] to test cross-domain robustness in a more realistic scenario, where the covariate shift is present due to differences between images from two different astronomical surveys. These differences are due to different levels of observational noise, PSF blurring, pixel scale, as well as differences in populations of observable astronomical objects (how distant or how faint a resolved object can be). GZ is a citizen science project that labels galaxy images through online participation. GZ Evo combines labeled image datasets across several surveys and iterations of GZ. Within GZ Evo, we use the GZ2 Dataset from the Sloan Digital Sky Survey (SDSS) [Willett et al., 2013] as the source, and a GZ Dark Energy Spectroscopic Instrument (DESI) dataset that combines observations from the DESI Imaging Surveys (DECals, MzLS, BASS, DES) [Walmsley et al., 2021, 2023] as the target. Older GZ SDSS data contains objects up to magnitude 17 in the r band, and redshifts below 0.25, while newer GZ DESI includes fainter objects up to magnitude 19 and more distant objects with redshifts below 0.4. Additionally, these surveys are different in the amount of observational noise and PSF blurring. Finally, GZ SDSS images include 3-filter gri images, while GZ DESI includes 3-filter grz images.

The original GZ Evo dataset contains $664,219$ images, each labeled according to vote counts (e.g., “8 of 10 volunteers answered Spiral, and 7 of 10 answered Bar”). GZ Evo also offers an aggregated version where the vote counts are converted into distinct classes, which is convenient for developing machine learning models. $239,408$ galaxies could be confidently assigned a distinct class based on the original vote counts. Of this sample, $82,185$ corresponded to GZ DESI and $95,703$ to GZ SDSS. The galaxy dataset used in this work contains a random sample of $40,000$ (32,000 training images with 20% set aside for validation, and 8,000 testing images) images in each of the domains, across six distinct classes: “smooth-round”, “smooth-cigar”, “unbarred spiral”, “edge-on-disk”, “barred spiral”, and “featured” galaxies. The “featured” galaxy class corresponds to any galaxy without a clear spiral structure or a visible bar, but which is also not completely smooth. The original galaxy images had dimensions $428 \times 428 \times 3$ and were subsequently downsampled to $100 \times 100 \times 3$ for more efficient training. Example images across all classes between GZ SDSS and GZ DESI are shown in Figure 3.

4. Network Architectures and Experiments

We evaluate our method on two sets of NNs: (1) CNNs constructed in `PyTorch` and (2) ENNs constructed in `escnn`, a PyTorch-based library for easy construction

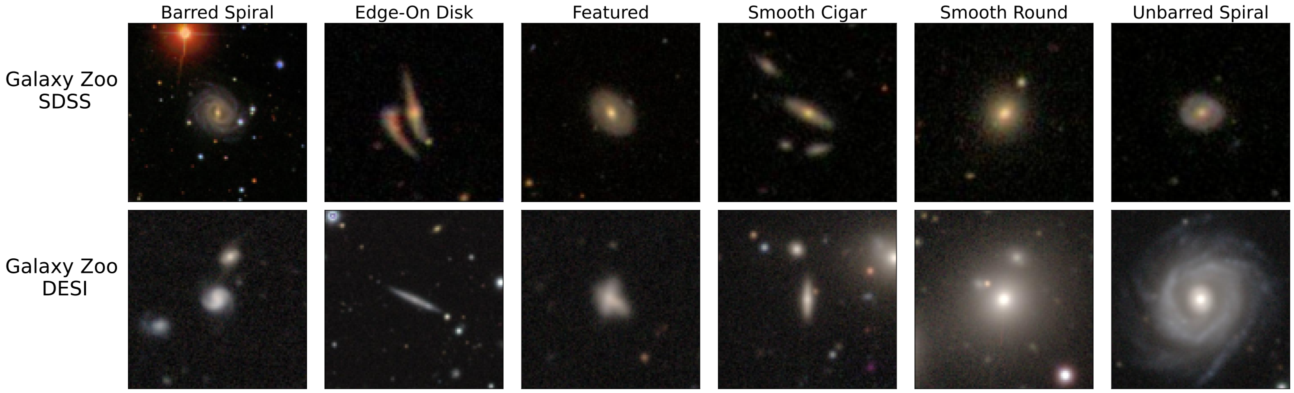

Figure 3. Top Panel: Example source domain images from the GZ Evo dataset with corresponding labels. Images are from GZ2 data observed by SDSS. **Bottom Panel:** Example target domain images from the GZ Evo dataset with the same labels. Images are from GZ DESI (combined observations from the DESI Imaging Surveys).

of ENNs. Details of our CNN and ENN architectures can be found in Appendix A. Most of our experiments use an ENN equivariant to D_4, following the results in our previous work [Pandya et al., 2023], as this level of equivariance is beneficial but not overly computationally expensive to train. We also study the performance of our method as a function of group order for the dihedral group. The code used in this work is available on our GitHub.

4.1. Equivariant Neural Networks

The efficacy of DA is limited by the feature extraction capabilities of the NN and its performance on the source domain. For image classification tasks, CNNs are natural choices due to their translation invariance and locality. There are, however, often additional symmetries inherent in the data, such as rotational and reflection invariance, that can be leveraged to enhance feature extraction and improve performance.

ENNs are a subclass of CNNs that can exploit higher-order symmetries besides the typical translation equivariance of CNNs [Cohen and Welling, 2016a,b]. Of interest for 2D images are symmetries of the Euclidean group $E(2)$ — in particular, the 2D special orthogonal group SO(2) and the orthogonal group O(2) and its associated subgroups. These (sub)groups allow ENNs to inherit symmetries of the circle and N-gon, respectively. As SO(2) and O(2) are continuous, they contain an infinite number of irreducible representations and have associated challenges when constructing architectures. For this reason, the discrete subgroup of O(2), the dihedral group D_N, is used in this work, which is straightforward to construct using open-source software, such as `escnn` [Cesa et al., 2022, Weiler and Cesa, 2019].

For image data (particularly astronomical data), rotational symmetry (with or without reflections) is typically inherent. For instance, galaxy morphologies are often invariant under rotations because there is no preferred reference frame in the Universe. Learning these symmetries can be induced in typical CNNs through data augmentation during training [Hernández-García and König, 2020, Wang et al., 2022b]. However, the output feature map f_{out} with grid values (ζ, ξ) from the convolution with kernel K with grid values (i, j),

$$f_{\text{out}}(\zeta, \xi) = \sum_{i,j} K(i, j) \cdot f(\zeta + i, \xi + j), \qquad (14)$$

is inherently only translation-equivariant. In contrast, group convolution [Cohen and Welling, 2016a, Kondor and Trivedi, 2018] in ENNs can be equivariant to any arbitrary group G:

$$f_{\text{out}}(g) = \sum_{h \in G} K(h^{-1}g) \cdot f(h) \qquad (15)$$

for $g, h \in G$: g and h correspond to the transformation for which the output feature map $f_{\text{out}}(g)$ is computed. For example, if G is the group of 2D rotations SO(2), g and h represent specific rotation angles. ENNs construct specialized filters K that are equivariant to the desired symmetries. Therefore, they learn transformation-invariant features that preserve the underlying symmetries in the data throughout training.

Each NN operation — convolution, activation, pooling, and dropout — must be equivariant [Srivastava et al., 2014]. Our ENN architectures have three convolutional blocks, each containing a group convolution (`R2Conv`), batch normalization (`InnerBatchNorm`), ReLU activation, max pooling (`MaxPoolPointwise2D`), and dropout (`PointwiseDropout`). The global group equivariance for the network is defined before the first convolutional layer. Following the convolutional layers, a group pooling operation aggregates over all the symmetry channels, an essential step for constructing invariant representations in the latent distribution as discussed in Hansen et al. [2024]. Finally, the ENN includes two linear layers that downsample the learned representation to the designated number of output classes for classification.

The latent vector of the NN is extracted before the last linear layer, with no dropout in the linear layers. The latent vectors for all our networks undergo a layer normalization [Ba et al., 2016], which serves to stabilize the latent distributions by standardizing the feature distributions across each sample, ensuring consistent scaling and preventing the activations from drifting to extreme values. We found that the layer norm is useful when employing SIDDA during training for more stable computations of σ_ℓ as given in Equation 7.

The architectures used in our experiments are based on the D_N group, which exhibits reflection symmetry. In the majority of our experiments, we set $N = 4$, as it provides notable benefits without imposing significant computational overhead during training. Our CNN has the same architecture, except the network components are not equivariant.

4.2. Training

We train all networks typically ($\mathcal{L} = \mathcal{L}_{\mathrm{CE}}$) and with SIDDA (Equation 6) and study performance differences in the source and target domain for the two techniques. In all experiments, we use the AdamW optimizer [Loshchilov and Hutter, 2019] with an initial learning rate of 10^{-2}, a weight decay of 10^{-3}, and a batch size of 128. A multiplicative learning rate decay of 0.1 is applied twice sequentially during training to stabilize convergence.

For all experiments, we use data augmentation comprising random rotations, flips, and affine translations. For the CNN, the augmentation instills approximate equivariance to encourage the model to learn rotation-invariant features. This is done to prevent predisposing the ENNs to perform better, but it also encourages faster convergence for the CNN. This later allows a more fruitful comparison between the latent distributions between ENNs, which have inherent rotation invariance to discrete rotations, and the CNNs, which have approximate invariance [Hansen et al., 2024].

For experiments with DA, an initial warm-up phase is implemented, during which only classification tasks are trained (using only $\mathcal{L}_{\mathrm{CE}}$). A similar 0.1 multiplicative learning rate decay as in the case of experiments without DA is also used. Early stopping and model-saving criteria are based on the following: for experiments without DA, we use the best validation loss on the source domain for classification; for DA experiments, we use the sum of the validation classification loss on the source domain and validation DA loss. The warm-up phase is intended to predispose the model to be at an ideal location in the loss landscape before starting DA. Empirically, we found this beneficial. The duration of the warm-up phase was tuned for each experiment, ensuring that it was long enough that the models were performant on the source domain but short enough that there was no overfitting. It was also found that ENNs required a shorter warmup phase than CNNs. For example, for the shapes dataset, a warm-up of five and 10 epochs were used for the D_4 and the CNN models, respectively. For the GZ Evo dataset, the warm-up phase was 20 and 30 epochs for the D_4 and the CNN models, respectively. For CNN experiments, models were trained for a maximum of 100 total epochs, while D_4 models were trained for a maximum of 60 total epochs, owing to the quicker convergence of ENNs.

All training was done on one NVIDIA A100-80GB GPU, with the most complex experiment requiring about one hour to train. A detailed algorithm describing one forward pass during training with SIDDA is given in Algorithm 1.

4.3. Calibration

Despite the impressive predictive capability of NNs in classification tasks across various fields, many real-world applications of NN-based classifiers also consider the confidence of each output class. Specifically, many NN-based classifiers can be uncalibrated, wherein the predicted class probabilities can frequently misrepresent the true class likelihood and lead to under or overconfident predictions. In data-sensitive or safety-sensitive settings—such as medicine and biology—proper model calibration is essential for deploying NN-based classifiers [Carse et al., 2022]. Similarly, in cosmology, simulation-based inference (SBI) pipelines that rely on trained classifiers must ensure proper calibration to guarantee that the inferred likelihood ratios or posterior probabilities are accurate and trustworthy [Cole et al., 2022].

Calibration techniques vary from regularization during training (either through architectural choices or additional loss terms) to post hoc methods that scale predicted probabilities (see Wang [2024] for a review). DA-based methods, however, have traditionally not been considered in the realm of regularization methods for model calibration. We will show that including DA with SIDDA not only improves accuracy (see Sections 5.1 and 5.2), but also calibration.

We evaluate model calibration using the Brier score and the Expected Calibration Error (ECE). The Brier score is the mean prediction error over all the classes:

$$\text{Brier Score} = \frac{1}{C}\sum_{i=1}^{C}(y_i - \delta_{i\hat{y}})^2, \tag{16}$$

where y_i is the NN-predicted score for each class $i \in C$, where C is the number of classes, $\hat{y}$ is the true class label, and $\delta_{i\hat{y}}$ is the Kronecker delta. Thus, a lower Brier Score is indicative of better calibration.

The ECE is the weighted average of the absolute difference between accuracy and confidence over V equally spaced confidence bins. For each bin B_v, where $v \in \{1, \ldots, V\}$, the accuracy and confidence are calculated based on the predictions within that bin. The ECE is defined as

$$\mathrm{ECE} = \sum_{v=1}^{V} \frac{|B_v|}{W} \left| \mathrm{acc}(B_v) - \mathrm{conf}(B_v) \right| , \qquad (17)$$

where $|B_v|$ is the number of samples in bin v, W is the total number of samples, $\mathrm{acc}(B_v)$ is the average accuracy in bin B_v, and $\mathrm{conf}(B_v)$ is the confidence (average estimated probability) in bin B_v. The ECE is thus a measure of how much model confidence aligns with the true distribution of classes, and a lower ECE indicates better calibration.

4.4. Neural Network Latent Distributions

The distributions over the source and target latent encodings, z and z^*, are the fundamental objects used in DA techniques. Probing the latent distributions can give crucial insights into the success and failure points of DA. The dimensionality of latent distributions is typically too large for visualization and analysis (256 in experiments used in this work). Therefore, dimensionality reduction techniques are often employed before visualizing the latent distributions. Techniques like t-SNE and UMAP [van der Maaten and Hinton, 2008, McInnes et al., 2018] use local metrics, like pairwise distances or nearest-neighbor graphs to preserve the structure of the data at small scales while embedding it into a lower-dimensional space. However, local metrics, and therefore these techniques, are limited because they primarily focus on preserving relationships within small neighborhoods of the data, often at the expense of capturing global structures or long-range dependencies that are critical for understanding the overall geometry or topology of the dataset. In contrast, the isomap [Tenenbaum et al., 2000] is a non-linear dimensionality reduction technique that estimates the global geometry of a latent vector manifold by using information of the nearest neighbors for each point in the latent space.

In this work, we use isomaps to visualize latent distributions, and we use the mean Silhouette score to quantify the inter-class (between clusters) and intra-class (within a cluster) distances and evaluate the quality of the clustering. The silhouette score is

$$s = \frac{1}{Q} \sum_{i=1}^{Q} \frac{b(i) - a(i)}{\max(a(i), b(i))}, \qquad (18)$$

where Q is the number of points; $a(i)$ is the intra-cluster distance, which is the mean distance between the i-th data point and all other points within the same cluster; and $b(i)$ is the inter-cluster distance, the mean distance between the i-th data point and points in the nearest neighboring cluster. The Silhouette score is in the range $[-1, 1]$, where values close to one indicate well-clustered data, values near zero indicate overlapping clusters, and negative values indicate that the data point may be assigned to the wrong cluster.

5. Results

All results are computed from three trained NNs, each with a different random seed for initializing weights. For each set of three trained networks, we estimate 1σ uncertainties on our diagnostic metrics. We refer to a model trained without DA (i.e., only with cross-entropy loss) as "<model>" — e.g., "CNN" or "D_4". In contrast, we refer to a model trained with SIDDA as "<model>-DA" — e.g., "CNN-DA" or "D_4-DA".

5.1. Simulated Datasets

The test set accuracies for the CNN, D_4, CNN-DA, and D_4-DA models on the shapes, astronomical objects, MNIST-M, and GZ Evo datasets are shown in Table 1.

For the shapes and simulated astronomical objects dataset, both models achieve near-perfect accuracy (above 99%) on the source domain without DA, but classification accuracy is much lower on the target domain for both (between 50% and 66%). With the inclusion of DA, the largest increase in target domain accuracy of ≈ 40% is for the astronomical objects datasets with CNN-DA. Despite this, the D_4 model significantly outperforms the CNN in the target domain, in the case of both datasets. With DA, the D_4-DA model achieves greater than 97% accuracy in both the source and target domains. While the CNN-DA shows a substantial improvement in target domain performance, the gap between the CNN-DA and D_4-DA remains considerable, with a difference of 21% in target domain accuracy for the shapes dataset.

We test generalization capabilities in the presence of Poisson noise and PSF blurring on the MNIST-M dataset. A similar trend emerges: the D_4 model is significantly more robust against both types of noise compared to the CNN. As in previous cases, with DA, neither model achieves the same performance as on the source domain (approximately 95% for CNN-DA and 97% for D_4-DA). However, the D_4-DA model demonstrates a greater potential for alignment than CNN-DA, achieving 93% accuracy in the target domain during the PSF blurring experiments.

The results show that SIDDA improves the source domain performance across most datasets for both CNN

Table 1. Classification accuracies for different model configurations on all datasets.

Dataset	Metric	CNN	CNN-DA	D_4	D_4-DA
Shapes	Source Acc. (%)	99.80 ± 0.04	$\mathbf{99.82 \pm 0.12}$	99.90 ± 0.04	$\mathbf{99.92 \pm 0.02}$
	Target Acc. (%)	50.47 ± 8.39	$\mathbf{78.20 \pm 1.73}$	64.76 ± 3.42	$\mathbf{99.71 \pm 0.06}$
Astro. Objects	Source Acc. (%)	$\mathbf{99.34 \pm 0.21}$	95.32 ± 1.57	$\mathbf{99.98 \pm 0.02}$	99.89 ± 0.50
	Target Acc. (%)	50.81 ± 2.89	$\mathbf{91.33 \pm 1.41}$	66.41 ± 2.07	$\mathbf{97.19 \pm 0.51}$
MNIST-M (Noise)	Source Acc. (%)	$\mathbf{95.64 \pm 0.12}$	95.31 ± 0.09	97.30 ± 0.30	$\mathbf{97.45 \pm 0.02}$
	Target Acc. (%)	68.32 ± 2.72	$\mathbf{76.24 \pm 1.12}$	70.31 ± 0.96	$\mathbf{87.55 \pm 0.16}$
MNIST-M (PSF)	Source Acc. (%)	95.64 ± 0.12	$\mathbf{95.66 \pm 0.13}$	97.30 ± 0.30	$\mathbf{97.95 \pm 0.10}$
	Target Acc. (%)	75.00 ± 1.44	$\mathbf{85.68 \pm 1.66}$	77.85 ± 1.31	$\mathbf{93.00 \pm 1.14}$
Galaxy Zoo Evo	Source Acc. (%)	81.49 ± 0.32	$\mathbf{81.57 \pm 0.82}$	86.65 ± 0.31	$\mathbf{87.58 \pm 0.06}$
	Target Acc. (%)	70.65 ± 2.26	$\mathbf{77.54 \pm 0.62}$	79.48 ± 1.52	$\mathbf{83.13 \pm 0.53}$

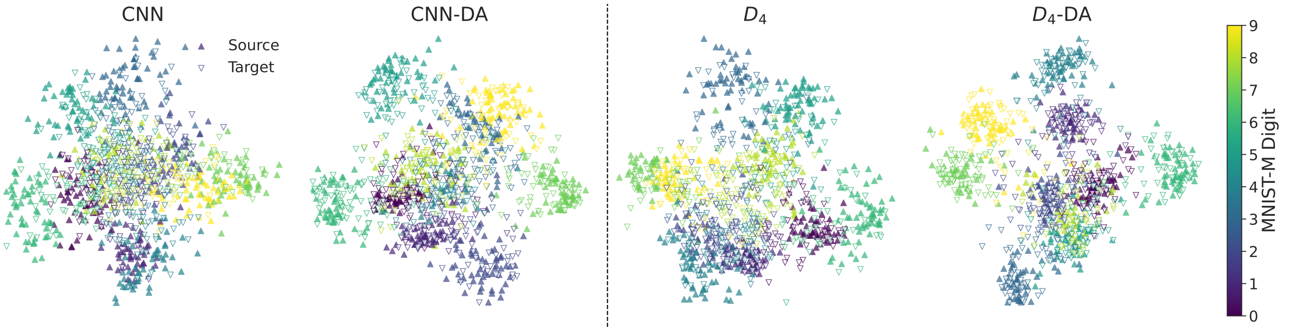

Figure 4. MNIST-M (Noise) latent distributions visualized with isomaps. Source (solid) and target latent (hollow) distributions are plotted atop each other to visualize latent distribution misalignment. The inclusion of DA clearly improves the alignment of source and target latent distributions for both CNN and D_4 models. It is also seen that the latent distribution of the D_4 is more clustered than the CNN, and even more so when D_4-DA and CNN-DA are compared. The improved clustering and separation of classes in the latent space is suggestive of improved feature learning.

and D_4 models, with the exception of the astronomical objects dataset for both models, and for the MNIST-M (noise) dataset for the CNN-DA model. In most cases, this drop in the source domain performance is below 1% (except in the case of astronomical objects classified with the CNN-DA model), which is acceptable given that the inclusion of DA substantially improves the accuracy on the unlabeled target domain. Additionally, the D_4-DA models converged to more similar performance across different seeds, as reflected by the small uncertainties in classification accuracy compared to CNN-DA.

We visualize the latent distributions of the CNN, D_4, CNN-DA, and D_4-DA in Figure 4 for the MNIST-M dataset with Poisson noise with the source (solid triangles) and target domain (hollow triangles) overlapped using isomaps [Tenenbaum et al., 2000]. Without DA, there is significant overlap between different classes, particularly in the middle of the figure for both models, though more apparent for the CNN. This is reflected in the CNN and D_4 Silhouette scores (Table 2), indicating similar levels of misclassified points, as seen in the target domain accuracy of approximately 68% for CNN and 70% for D_4. With the inclusion of DA, there is better alignment of the same classes from source and target latent distributions for both models. Furthermore, with DA, there is increased class separation, especially along the periphery, but there is still some overlap in the middle of the diagram (particularly for CNN-DA). Compared to the CNN latent distribution, the class separation for the D_4 latent distribution is much larger due to the equivariance in the D_4 model, which is reflected in the target domain Silhouette scores increasing more significantly for the D_4-DA when compared to the CNN-DA.

Model	Source	Target
CNN	0.1930	-0.0539
CNN-DA	0.3360	0.1150
D_4	0.2744	-0.0247
D_4-DA	**0.4023**	**0.1983**

Table 2. Silhouette scores for CNN, D_4, CNN-DA, and D_4-DA on MNIST-M (Noise).

5.2. Galaxy Zoo Evo Dataset

We aim to address a common problem in real astronomical applications: the scenario where one has access to lower-quality, older observations with labels, but no labels exist for the newer, higher-quality target dataset from a more recent astronomical survey. To test our model in this situation, we use galaxy datasets from GZ Evo [Walmsley et al., 2024]. Namely, we use GZ2 from SDSS as the source domain and GZ DESI as the target domain. This dataset is considerably larger than the previous simulated datasets. The results for these experiments are summarized in Table 1. Similar to the results for the simulated datasets, the D_4 model fully outperforms the CNN, with a $\approx 9\%$ higher accuracy in the target domain. With DA, both CNN-DA and D_4-DA models have higher accuracy in both the source and target domain. The performance difference between the two kinds of models is more moderate for the GZ Evo dataset than for the simulated datasets. This may be attributable to the GZ Evo data's larger size, greater morphological complexity, and the use of data augmentation during training. In the large data limit, the inclusion of data augmentation causes CNNs to become approximately equivariant [Hernández-García and König, 2020, Wang et al., 2022b]. Nevertheless, the D_4-DA achieves $\sim 6\%$ higher accuracy in the target domain compared to the CNN-DA model.

The accuracy of all models in this experiment was lower than the experiments with the simulated data, with no model achieving greater than $\approx 88\%$ accuracy in either the source or target domain. This dataset is significantly larger than the others, so a deeper or larger model with additional training aids, such as residual connections, would likely yield better performance [Nakkiran et al., 2019, He et al., 2015]. Our goal is to study the efficacy of SIDDA, and not to achieve state-of-the-art performance on this dataset, so we did not experiment with a more complex model.

5.3. Robustness with Group Order

Table 3. Performance results for different orders of the dihedral group D_N, without and with DA.

Group	Source Domain	Target Domain
D_1	$96.03 \pm 0.18\%$	$63.69 \pm 0.61\%$
D_2	$97.06 \pm 0.048\%$	$67.88 \pm 2.5\%$
D_4	$97.30 \pm 0.28\%$	$70.30 \pm 0.97\%$
D_8	$97.42 \pm 0.10\%$	$71.67 \pm 0.28\%$
D_1-DA	$95.40 \pm 0.084\%$	$75.35 \pm 0.71\%$
D_2-DA	$97.10 \pm 0.23\%$	$84.98 \pm 1.9\%$
D_4-DA	$97.50 \pm 0.074\%$	$87.70 \pm 0.26\%$
D_8-DA	$\mathbf{97.69 \pm 0.081\%}$	$\mathbf{88.96 \pm 0.32\%}$

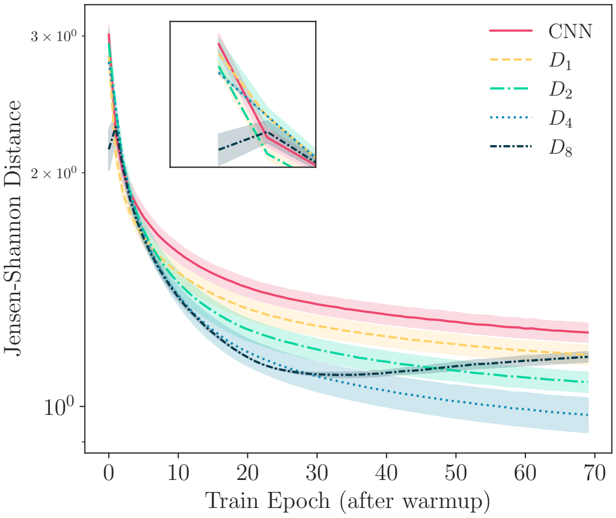

Figure 5. Jensen-Shannon (JS) distances for CNN-DA and D_N-DA ($N \in \{1, 2, 4, 8\}$) models trained on MNIST-M (Noise). Shaded regions correspond to 1σ uncertainties from three training runs initialized with varying random seeds. All models underwent a 30-epoch warm-up phase without DA, and the JS distance is shown for epochs thereafter, which is where SIDDA is used. Compared to the CNN, all D_N models exhibit a lower JS distance between source and target domains after the warm-up phase, which can be attributed to the fact that the equivariance constraint encourages distributional similarity between the source and target latent distributions. We see that the D_N models also achieve more perfect alignment with the introduction of DA, as shown by the lower JS distances by the end of the training. This behavior correlates with the group order, except for D_8-DA, which achieved its best model much earlier in training and began overfitting.

Next, we study the robustness of ENNs, both with and without SIDDA, and with increasing orders of the dihedral group D_N, with $N \in \{1, 2, 4, 8\}$ on MNIST-M with Poisson noise in the target domain. The robustness of ENNs as a function of group order was previously studied in [Pandya et al., 2023] for generalization to Poisson noise in images in the task of galaxy morphology classification. That work showed that robustness generally increased with group order, but high group-order models tended to overfit or have lower accuracy as a result of the equivariance constraint becoming too strong (see Figure 2 in Pandya et al. [2023]). Additionally, despite the robustness, there was no significant overlap in the latent distribution of the NNs when introduced to covariate shifts in the data. In the current work, we add to that previous study by examining the effect of SIDDA on the robustness of ENNs across different group orders with the same data setup.

We follow a training procedure similar to that outlined in Section 4.2 for all networks. The major difference is that we do not implement early stopping so that we can study the propensity for overfitting as a function of group order. Models are still saved based on the best validation loss, and we show results from the best-performing models on the held-out test set. Source and target domain classification accuracies are shown in Table 3. Both source and target accuracies increase with group order, with and without DA. There is also a slight increase in source domain accuracy when using DA, except for D_1, which contains a single reflection and therefore exhibits trivial equivariance.

We also track the evolution of the mean JS distance between source and target latent distributions for all models during training (after the initial warm-up phase has passed). Assuming perfect feature learning and optimal performance on the source domain, minimizing the JS distance between source and target domain latent distributions corresponds to optimal performance on the target domain [Shui et al., 2022]. As shown in Figure 5, after the initial 30-epoch warm-up phase, the JS distance decreases as the group order increases. Among the models, the D_8-DA model exhibits the lowest JS distance at the end of the warm-up, while the CNN-DA exhibits the highest. This supports the claim that the constrained latent distribution of ENNs leads to significantly better alignment between source and target domains, as made in Section 2.4.

As training progresses with DA, the JS distance generally decreases with increasing group order, except for the D_8 model. This model begins to overfit 30 epochs after the warm-up phase concludes, as indicated by the JS distance in Figure 5. This was also confirmed upon inspection of the validation loss. All other models reach their highest validation loss after epoch 90. Despite this overfitting, the best D_8-DA model still achieves the highest source and target domain accuracy, as shown in Table 3.

The overfitting observed in the D_8 model can be attributed to the stronger equivariance constraint (i.e., more weight sharing), which may limit the model's expressivity when the covariate shifts in the target domain do not fully respect the underlying symmetry. That is, the allowable space of features that respects the stronger equivariance will be inherently smaller than those respected by more lenient equivariance or the typical translation equivariance in CNNs, considering that ENNs assume perfect symmetries in the data. In the presence of perturbations, which is a typical case in the target domain for DA applications, this rarely holds. Solutions to relaxing the equivariance constraint while still enjoying the benefits of symmetry constraints have been extensively studied in other works [Wang et al., 2022a, Elsayed et al., 2020].

5.4. Calibration Results

For all datasets, we observe that the D_4 and D_4-DA models result in lower ECE and Brier scores across most experiments in both the source and target domain when compared to the CNN and CNN-DA. Across all experiments, the largest improvement is observed for the D_4-DA model applied to the shapes target domain data, where both the ECE and Brier score are reduced by more than an order of magnitude compared to the regular D_4 model.

With the inclusion of DA, there is an improvement in the calibration scores for both data domains and for all models for the MNIST-M datasets, as well as GZ Evo data. Between these datasets, the largest improvement is for the D_4-DA model on the MNIST-M (PSF) dataset in the target domain, exhibiting an approximate factor of two reduction for both the ECE and Brier score. However, at the same time, we observe a decrease in the source domain calibration for all DA experiments in the shapes and astronomical objects datasets, with the exception of D_4-DA in the shapes dataset. This indicates that in the source domain, at least for more simple datasets, the inclusion of DA can potentially worsen the calibrations of the models. This decrease in performance is not always apparent in the (uncalibrated) accuracies, as shown in Table 1. For instance, the source accuracies with and without DA are very similar (within the margin of error) in all cases, except for the CNN-DA model applied to the astronomical objects dataset. Nonetheless, in the target domain, we observe strict improvements in both accuracy and calibration across all experiments with the inclusion of DA with SIDDA, as shown in Table 1 and Table 4.

6. Conclusions and Outlook

In this work, we introduced SIDDA — a method for semi-supervised DA that includes principled methods for optimal alignment of NN latent spaces and training with multiple loss terms. This is in contrast with most DA applications, which face the challenge of requiring extensive hyperparameter tuning that makes training NNs time-consuming, expensive, or unfeasible. SIDDA is an “out-of-the-box” DA method that is employable in various domains, and it requires labeled training data only in the source domain, not in the target domain.

Our method relies on the Sinkhorn divergence, a symmetrized variant of regularized OT distances, which corrects for the bias that exists in OT_σ. In particular, our method dynamically adjusts the strength of regularization on a per-epoch basis dependent on the pairwise distance between the source and target latent distribution vectors. In addition, we use dynamic weighting of the cross-entropy and DA loss terms on a per-epoch basis, which ensures a balance in training

Table 4. Calibration metrics for different model configurations.

Metric	CNN	CNN-DA	D_4	D_4-DA
Shapes				
Source ECE	**0.011 ± 0.001**	0.013 ± 0.001	0.011 ± 0.002	**0.0074 ± 0.0003**
Source Brier	**0.000734 ± 0.000090**	0.00112 ± 0.00020	0.000814 ± 0.000200	**0.000349 ± 0.000034**
Target ECE	0.35 ± 0.04	**0.29 ± 0.004**	0.20 ± 0.03	**0.013 ± 0.002**
Target Brier	0.110 ± 0.010	**0.0925 ± 0.002**	0.0564 ± 0.009	**0.0015 ± 0.0003**
Astronomical Objects				
Source ECE	**0.041 ± 0.010**	0.075 ± 0.020	**0.00695 ± 0.00030**	0.00899 ± 0.00090
Source Brier	**0.00798 ± 0.003**	0.0220 ± 0.006	**0.000132 ± 0.000031**	0.000746 ± 0.000300
Target ECE	0.17 ± 0.04	**0.142 ± 0.010**	0.294 ± 0.020	**0.053 ± 0.008**
Target Brier	0.0440 ± 0.010	**0.0420 ± 0.006**	0.0804 ± 0.009	**0.0150 ± 0.003**
MNIST-M (Noise)				
Source ECE	0.161 ± 0.003	**0.126 ± 0.004**	0.114 ± 0.002	**0.0790 ± 0.002**
Source Brier	0.00991 ± 0.00023	**0.00880 ± 0.00030**	0.00610 ± 0.00030	**0.00481 ± 0.00010**
Target ECE	0.409 ± 0.013	**0.355 ± 0.009**	0.390 ± 0.020	**0.250 ± 0.009**
Target Brier	0.0450 ± 0.003	**0.0370 ± 0.002**	0.0410 ± 0.0008	**0.0210 ± 0.0031**
MNIST-M (PSF)				
Source ECE	0.161 ± 0.003	**0.124 ± 0.004**	0.114 ± 0.002	**0.0750 ± 0.002**
Source Brier	0.00991 ± 0.00023	**0.00850 ± 0.00040**	0.00610 ± 0.00030	**0.00400 ± 0.00020**
Target ECE	0.384 ± 0.013	**0.272 ± 0.012**	0.340 ± 0.020	**0.181 ± 0.001**
Target Brier	0.0340 ± 0.001	**0.0230 ± 0.001**	0.0270 ± 0.003	**0.0130 ± 0.00007**
Galaxy Zoo Evo				
Source ECE	0.283 ± 0.0019	**0.264 ± 0.0044**	0.2322 ± 0.00086	**0.206 ± 0.0011**
Source Brier	0.0453 ± 0.00031	**0.0439 ± 0.0010**	0.0341 ± 0.00059	**0.0319 ± 0.00014**
Target ECE	0.324 ± 0.0078	**0.301 ± 0.0018**	0.271 ± 0.0042	**0.241 ± 0.0026**
Target Brier	0.0538 ± 0.0015	**0.051 ± 0.00029**	0.0411 ± 0.0012	**0.0382 ± 0.00048**

and improves predictive performance in both the source and target domains.

We test our method on shapes and astronomical objects datasets simulated using `DeepBench`, the MNIST-M dataset, and a dataset of real galaxy images from Galaxy Zoo. These experiments encompass covariate shifts induced by Poisson noise, PSF blurring, and more complex differences between real astronomical surveys.

We draw the following conclusions about SIDDA:

- SIDDA requires minimal hyperparameter tuning to achieve a considerable increase in target domain performance;
- SIDDA is compatible with a wide range of models, including CNNs and ENNs equivariant to various groups. Its efficacy is more pronounced when paired with ENNs, offering in some cases nearly 50% better target domain accuracy when compared to CNNs trained without DA (Table 1);
- We find that SIDDA is effective across ENNs with varying group-order equivariance, and its performance improves as the degree of equivariance increases (Table 3 and Figure 5);
- SIDDA can improve the source and target domain performance of NNs, and its benefits are concretely seen when analyzing the clustering and alignment of NN latent spaces (Table 2 and Figure 4);
- Though not constructed as such, SIDDA can inherently improve the calibration of trained NN-based classifiers (Table 4);
- SIDDA does not stack any considerable computational expense when training models, as it builds upon existing, efficient coding frameworks. All models in this work were trained on one GPU in typically less than an hour.

There are multiple opportunities for further development of SIDDA. First, the metric used for adjusting the dynamic Sinkhorn plan S_σ relies on the pairwise norm between entries in the latent space. Other notions of distance to adjust the regularization of S_σ can be considered. Second, in this work, we implemented a manual truncation of η_i terms in the loss function (Equation 6) to ensure that the DA loss does not overpower the classification loss. Other levels of truncation or regularization of the loss should be studied to find the most optimal balance of the two loss terms. Third, all experiments performed here are for a fixed architecture, which employs many features of NNs that are now standard (dropout, pooling, batch normalization). Notably missing are residual connections, which have been found to aid the convergence of many NNs. It would be interesting to study the efficacy of SIDDA with deeper, more complex networks as well as ENNs equivariant to the continuous analogous of groups studied here (i.e., O(2)). Lastly, our method works with fixed classes and cannot operate when the classes between the source and target domain are not the same. A potential extension of SIDDA could involve making it compatible with a flexible number of classes in both the source and target domains, drawing inspiration from the DeepAstroUDA method [Ćiprijanović et al., 2023].

The problem of generalization in classification tasks in NNs can be primarily considered as a problem of robust feature learning (e.g., architectural choice) and domain alignment. The most successful domain adaptation methods should leverage principled choices for both aspects. ENNs are natural candidates for robust feature learning because their feature learning capabilities can be inherently constrained to symmetries of the data. However, most existing methods for domain adaptation implicitly require many empirical choices or hyperparameter tuning. In this work, we have combined these aspects in introducing SIDDA, which leverages a dynamic parameterization for OT-based DA hyperparameters during training, and works particularly well when paired with ENNs. Our future work will be in refining this approach, and further developing more automated DA algorithms.

Funding

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who've facilitated an environment of open discussion, idea generation, and collaboration. This community was important for the development of this project.

This work is supported by: the Fermi National Accelerator Laboratory, managed and operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy (DOE), the Office of Science; the Office of Workforce Development for Teachers and Scientists, the Office of Science Graduate Student Research (SCGSR) program, and the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). The SCGSR program is administered by the Oak Ridge Institute for Science and Education for the DOE under contract number DE-SC0014664. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

Acknowledgments and Author Contributions

The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

The following authors contributed in different ways to the manuscript.

Pandya: Conceptualization, Methodology, Investigation, Formal analysis, Software, Validation, Data curation, Resources, Visualization, Writing - Original Draft, Writing - Review & Editing, Project administration, Funding acquisition.

Patel: Methodology, Writing - Original Draft, Writing - Review & Editing.

Nord: Methodology, Writing - Original Draft, Writing - Review & Editing.

Walmsley: Data Curation, Writing - Review & Editing

Ćiprijanović: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

References

Shrihan Agarwal, Aleksandra Ćiprijanović, and Brian D. Nord. Neural Network Prediction of Strong Lensing Systems with Domain Adaptation and Uncertainty Quantification. *arXiv e-prints*, art. arXiv:2411.03334, October 2024. doi: 10.48550/arXiv.2411.03334.

Rachel Akeson, Lee Armus, Etienne Bachelet, Vanessa Bailey, Lisa Bartusek, Andrea Bellini, Dominic Benford, David Bennett, Aparna Bhattacharya, Ralph Bohlin, Martha Boyer, Valerio Bozza, Geoffrey Bryden, Sebastiano Calchi Novati, Kenneth Carpenter, Stefano Casertano, Ami Choi, David Content, Pratika Dayal, Alan Dressler, Olivier Doré, S. Michael Fall, Xiaohui Fan, Xiao Fang, Alexei Filippenko, Steven Finkelstein, Ryan Foley, Steven Furlanetto, Jason Kalirai, B. Scott Gaudi, Karoline Gilbert, Julien Girard, Kevin Grady, Jenny Greene, Puragra Guhathakurta, Chen Heinrich, Shoubaneh Hemmati, David Hendel, Calen Henderson, Thomas

Henning, Christopher Hirata, Shirley Ho, Eric Huff, Anne Hutter, Rolf Jansen, Saurabh Jha, Samson Johnson, David Jones, Jeremy Kasdin, Patrick Kelly, Robert Kirshner, Anton Koekemoer, Jeffrey Kruk, Nikole Lewis, Bruce Macintosh, Piero Madau, Sangeeta Malhotra, Kaisey Mandel, Elena Massara, Daniel Masters, Julie McEnery, Kristen McQuinn, Peter Melchior, Mark Melton, Bertrand Mennesson, Molly Peeples, Matthew Penny, Saul Perlmutter, Alice Pisani, Andrés Plazas, Radek Poleski, Marc Postman, Clément Ranc, Bernard Rauscher, Armin Rest, Aki Roberge, Brant Robertson, Steven Rodney, James Rhoads, Jason Rhodes, Russell Ryan Jr. au2, Kailash Sahu, David Sand, Dan Scolnic, Anil Seth, Yossi Shvartzvald, Karelle Siellez, Arfon Smith, David Spergel, Keivan Stassun, Rachel Street, Louis-Gregory Strolger, Alexander Szalay, John Trauger, M. A. Troxel, Margaret Turnbull, Roeland van der Marel, Anja von der Linden, Yun Wang, David Weinberg, Benjamin Williams, Rogier Windhorst, Edward Wollack, Hao-Yi Wu, Jennifer Yee, and Neil Zimmerman. The wide field infrared survey telescope: 100 hubbles for the 2020s, 2019. URL https://arxiv.org/abs/1902.05569.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, 2018. URL https://arxiv.org/abs/1705.09634.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.161. URL http://dx.doi.org/10.1109/TPAMI.2010.161.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Alexander Bogatskiy, Brandon Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics, 2020. URL https://arxiv.org/abs/2006.04780.

Srinath Bulusu, Matteo Favoni, Andreas Ipp, David I. Müller, and Daniel Schuh. Equivariance and generalization in neural networks. *EPJ Web of Conferences*, 258:09001, 2022. ISSN 2100-014X. doi: 10.1051/epjconf/202225809001. URL http://dx.doi.org/10.1051/epjconf/202225809001.

Jacob Carse, Andres Alvarez Olmo, and Stephen McKenna. Calibration of deep medical image classifiers: An empirical comparison using dermatology and histopathology datasets. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, page 89–99, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-16748-5. doi: 10.1007/978-3-031-16749-2_9. URL https://doi.org/10.1007/978-3-031-16749-2_9.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WE4qe9xlnQw.

A. Ćiprijanović, A. Lewis, K. Pedro, S. Madireddy, B. Nord, G. N. Perdue, and S. M. Wild. DeepAstroUDA: semi-supervised universal domain adaptation for cross-survey galaxy morphology classification and anomaly detection. *Machine Learning: Science and Technology*, 4(2):025013, June 2023. doi: 10.1088/2632-2153/acca5f.

Aleksandra Ćiprijanović, Diana Kafkes, Gregory Snyder, F. Javier Sánchez, Gabriel Nathan Perdue, Kevin Pedro, Brian Nord, Sandeep Madireddy, and Stefan M. Wild. DeepAdversaries: examining the robustness of deep learning models for galaxy morphology classification. *Machine Learning: Science and Technology*, 3(3):035007, September 2022. doi: 10.1088/2632-2153/ac7f1a.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks, 2016a. URL https://arxiv.org/abs/1602.07576.

Taco S. Cohen and Max Welling. Steerable cnns, 2016b. URL https://arxiv.org/abs/1612.08498.

Alex Cole, Benjamin K. Miller, Samuel J. Witte, Maxwell X. Cai, Meiert W. Grootes, Francesco Nattino, and Christoph Weniger. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *Journal of Cosmology and Astroparticle Physics*, 2022(09):004, September 2022. ISSN 1475-7516. doi: 10.1088/1475-7516/2022/09/004. URL http://dx.doi.org/10.1088/1475-7516/2022/09/004.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1853–1865, 2014. URL https://api.semanticscholar.org/CorpusID:13347901.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey, 2017. URL https://arxiv.org/abs/1702.05374.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas Guibas. Vector neurons: A general framework for so(3)-equivariant networks, 2021. URL https://arxiv.org/abs/2104.12229.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance, 2018. URL https://arxiv.org/abs/1610.06447.

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks, 2016. URL https://arxiv.org/abs/1604.04004.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions, 2017. URL https://arxiv.org/abs/1705.02498.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Bin Shao, and Tie-Yan Liu. Se(3) equivariant graph neural networks with complete local frames, 2022. URL https://arxiv.org/abs/2110.14811.

Gamaleldin F. Elsayed, Prajit Ramachandran, Jonathon Shlens, and Simon Kornblith. Revisiting spatial invariance with low-rank local connectivity, 2020. URL https://arxiv.org/abs/2002.02959.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns, 2018. URL https://arxiv.org/abs/1711.06721.

Abolfazl Farahani, Sahar Voghoei, Khaled M. Rasheed, and Hamid Reza Arabnia. A brief review of domain adaptation. *ArXiv*, abs/2010.03978, 2020. URL https://api.semanticscholar.org/CorpusID:222209143.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018. URL https://arxiv.org/abs/1810.08278.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Jean Feydy, Joan Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *Advances in Neural Information Processing Systems*, 33, 2020.

Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise, 2019. URL https://arxiv.org/abs/1901.10513.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020. URL https://arxiv.org/abs/2006.10503.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016. URL https://arxiv.org/abs/1505.07818.

Milind S. Gide, Samuel F. Dodge, and Lina J. Karam. The effect of distortions on the prediction of visual attention, 2016. URL https://arxiv.org/abs/1604.03882.

Sankalp Gilda, Antoine de Mathelin, Sabine Bellstedt, and Guillaume Richard. Unsupervised Domain Adaptation for Constraining Star Formation Histories. *Astronomy*, 3(3):189–207, July 2024. doi: 10.3390/astronomy3030012.

Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A kernel method for the two-sample problem, 2008. URL https://arxiv.org/abs/0805.2368.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Andreas Abildtrup Hansen, Anna Calissano, and Aasa Feragen. Interpreting equivariant representations, 2024. URL https://arxiv.org/abs/2401.12588.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization, 2020. URL https://arxiv.org/abs/1806.03852.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang, Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, March 2019. ISSN 1538-4357. doi: 10.3847/1538-4357/ab042c. URL http://dx.doi.org/10.3847/1538-4357/ab042c.

N. Jeffrey, L. Whiteway, M. Gatti, J. Williamson, J. Alsing, A. Porredon, J. Prat, C. Doux, B. Jain, C. Chang, T. Y. Cheng, T. Kacprzak, P. Lemos, A. Alarcon, A. Amon, K. Bechtol, M. R. Becker, G. M. Bernstein, A. Campos, A. Carnero Rosell, R. Chen, A. Choi, J. DeRose, A. Drlica-Wagner,

K. Eckert, S. Everett, A. Ferté, D. Gruen, R. A. Gruendl, K. Herner, M. Jarvis, J. McCullough, J. Myles, A. Navarro-Alsina, S. Pandey, M. Raveri, R. P. Rollins, E. S. Rykoff, C. Sánchez, L. F. Secco, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. A. Troxel, I. Tutusaus, T. N. Varga, B. Yanny, B. Yin, J. Zuntz, M. Aguena, S. S. Allam, O. Alves, D. Bacon, S. Bocquet, D. Brooks, L. N. da Costa, T. M. Davis, J. De Vicente, S. Desai, H. T. Diehl, I. Ferrero, J. Frieman, J. García-Bellido, E. Gaztanaga, G. Giannini, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. Huterer, D. J. James, O. Lahav, S. Lee, J. L. Marshall, J. Mena-Fernández, R. Miquel, A. Pieres, A. A. Plazas Malagón, A. Roodman, M. Sako, E. Sanchez, D. Sanchez Cid, M. Smith, E. Suchyta, M. E. C. Swanson, G. Tarle, D. L. Tucker, N. Weaverdyck, J. Weller, P. Wiseman, and M. Yamamoto. Dark energy survey year 3 results: likelihood-free, simulation-based $w$cdm inference with neural compression of weak-lensing map statistics, 2024. URL https://arxiv.org/abs/2403.02314.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Ioannis Kalogeropoulos, Giorgos Bouritsas, and Yannis Panagakis. Scale equivariant graph metanetworks, 2024. URL https://arxiv.org/abs/2406.10685.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, 2019. URL https://api.semanticscholar.org/CorpusID:57572938.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018. URL https://arxiv.org/abs/1705.07115.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups, 2018. URL https://arxiv.org/abs/1802.03690.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214%2Faoms%2F1177729694.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives. *arXiv e-prints*, art. arXiv:2208.07422, August 2022. doi: 10.48550/arXiv.2208.07422.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10 (1–2):1–141, 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt, 2019. URL https://arxiv.org/abs/1912.02292.

Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, Luke Kelley, Mark Lovell, Benedikt Diemer, and Lars Hernquist. The illustristng simulations: Public data release, 2021. URL https://arxiv.org/abs/1812.05609.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 2018. URL https://api.semanticscholar.org/CorpusID:88523547.

Sneh Pandya, Purvik Patel, Franc O, and Jonathan Blazek. E(2) equivariant neural networks for robust galaxy morphology classification, 2023. URL https://arxiv.org/abs/2311.01500.

Hanna Parul, Sergei Gleyzer, Pranath Reddy, and Michael W. Toomey. Domain adaptation in application to gravitational lens finding. *arXiv e-prints*, art. arXiv:2410.01203, October 2024. doi: 10.48550/arXiv.2410.01203.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Sashank J. Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions, 2014. URL https://arxiv.org/abs/1406.2083.

Andrea Roncoli, Aleksandra Ćiprijanović, Maggie Voetberg, Francisco Villaescusa-Navarro, and Brian Nord. Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets. *arXiv e-prints*, art. arXiv:2311.01588, November 2023. doi: 10.48550/arXiv.2311.01588.

Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.

Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density, 2021. URL https://arxiv.org/abs/2108.10860.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

R. Scaramella, J. Amiaux, Y. Mellier, C. Burigana, C. S. Carvalho, J.-C. Cuillandre, A. Da Silva, A. Derosa, J. Dinis, E. Maiorano, M. Maris, I. Tereno, R. Laureijs, T. Boenke, G. Buenadicha, X. Dupac, L. M. Gaspar Venancio, P. Gómez-Álvarez, J. Hoar, J. Lorenzo Alvarez, G. D. Racca, G. Saavedra-Criado, J. Schwartz, R. Vavrek, M. Schirmer, H. Aussel, R. Azzollini, V. F. Cardone, M. Cropper, A. Ealet, B. Garilli, W. Gillard, B. R. Granett, L. Guzzo, H. Hoekstra, K. Jahnke, T. Kitching, T. Maciaszek, M. Meneghetti, L. Miller, R. Nakajima, S. M. Niemi, F. Pasian, W. J. Percival, S. Pottinger, M. Sauvage, M. Scodeggio, S. Wachter, A. Zacchei, N. Aghanim, A. Amara, T. Auphan, N. Auricchio, S. Awan, A. Balestra, R. Bender, C. Bodendorf, D. Bonino, E. Branchini, S. Brau-Nogue, M. Brescia, G. P. Candini, V. Capobianco, C. Carbone, R. G. Carlberg, J. Carretero, R. Casas, F. J. Castander, M. Castellano, S. Cavuoti, A. Cimatti, R. Cledassou, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, A. Costille, F. Courbin, H. Degaudenzi, M. Douspis, F. Dubath, C. A. J. Duncan, S. Dusini, S. Farrens, S. Ferriol, P. Fosalba, N. Fourmanoit, M. Frailis, E. Franceschi, P. Franzetti, M. Fumana, B. Gillis, C. Giocoli, A. Grazian, F. Grupp, S. V. H. Haugan, W. Holmes, F. Hormuth, P. Hudelot, S. Kermiche, A. Kiessling, M. Kilbinger, R. Kohley, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, O. Lahav, S. Ligori, P. B. Lilje, I. Lloro, O. Mansutti, O. Marggraf, K. Markovic, F. Marulli, R. Massey, S. Maurogordato, M. Melchior, E. Merlin, G. Meylan, J. J. Mohr, M. Moresco, B. Morin, L. Moscardini, E. Munari, R. C. Nichol, C. Padilla, S. Paltani, J. Peacock, K. Pedersen, V. Pettorino, S. Pires, M. Poncet, L. Popa, L. Pozzetti, F. Raison, R. Rebolo, J. Rhodes, H.-W. Rix, M. Roncarelli, E. Rossetti, R. Saglia, P. Schneider, T. Schrabback, A. Secroun, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, J. Skottfelt, L. Stanco, J. L. Starck, P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, H. I. Teplitz, R. Toledo-Moreo, F. Torradeflot, M. Trifoglio, E. A. Valentijn, L. Valenziano, G. A. Verdoes Kleijn, Y. Wang, N. Welikala, J. Weller, M. Wetzstein, G. Zamorani, J. Zoubian, S. Andreon, M. Baldi, S. Bardelli, A. Boucaud, S. Camera, D. Di Ferdinando, G. Fabbian, R. Farinelli, S. Galeotta, J. Graciá-Carpio, D. Maino, E. Medinaceli, S. Mei, C. Neissner, G. Polenta, A. Renzi, E. Romelli, C. Rosset, F. Sureau, M. Tenti, T. Vassallo, E. Zucca, C. Baccigalupi, A. Balaguera-Antolínez, P. Battaglia, A. Biviano, S. Borgani, E. Bozzo, R. Cabanac, A. Cappi, S. Casas, G. Castignani, C. Colodro-Conde, J. Coupon, H. M. Courtois, J. Cuby, S. de la Torre, S. Desai, H. Dole, M. Fabricius, M. Farina, P. G. Ferreira, F. Finelli, P. Flose-Reimberg, S. Fotopoulou, K. Ganga, G. Gozaliasl, I. M. Hook, E. Keihanen, C. C. Kirkpatrick, P. Liebing, V. Lindholm, G. Mainetti, M. Martinelli, N. Martinet, M. Maturi, H. J. McCracken, R. B. Metcalf, G. Morgante, J. Nightingale, A. Nucita, L. Patrizii, D. Potter, G. Riccio, A. G. Sánchez, D. Sapone, J. A. Schewtschenko, M. Schultheis, V. Scottez, R. Teyssier, I. Tutusaus, J. Valiviita, M. Viel, W. Vriend, and L. Whittaker. Euclid preparation: I. the euclid wide survey. *Astronomy &; Astrophysics*, 662:A112, June 2022. ISSN 1432-0746. doi: 10.1051/0004-6361/202141938. URL http://dx.doi.org/10.1051/0004-6361/202141938.

J. L. Sersic. Photometry of southern galaxies: NGC 5128. *The Observatory*, 78:24–29, February 1958.

Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. A novel domain adaptation theory with jensen–shannon divergence. *Knowledge-Based Systems*, 257:109808, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109808. URL https://www.sciencedirect.com/science/article/pii/S0950705122009200.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, October 2019. ISSN 1941-0026. doi: 10.1109/tevc.2019.2890858. URL http://dx.doi.org/10.1109/TEVC.2019.2890858.

Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *ArXiv*, abs/1612.01939, 2016. URL https://api.semanticscholar.org/CorpusID:10084602.

Paxson Swierc, Megan Zhao, Aleksandra Ćiprijanović, and Brian Nord. Domain Adaptation for Measurements of Strong Gravitational Lenses. *arXiv e-prints*, art. arXiv:2311.17238, November 2023. doi: 10.48550/arXiv.2311.17238.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500):2319–2323, December 2000. ISSN 1095-9203. doi: 10.1126/science.290.5500.2319. URL http://dx.doi.org/10.1126/science.290.5500.2319.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ricardo Vilalta, Kinjal Dhar Gupta, Dainis Boumber, and Mikhail M. Meskhi. A general approach to domain adaptation with applications in astronomy. *Publications of the Astronomical Society of the Pacific*, 131(1004):108008, September 2019. ISSN 1538-3873. doi: 10.1088/1538-3873/aaf1fc. URL http://dx.doi.org/10.1088/1538-3873/aaf1fc.

Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, July 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abf7ba. URL http://dx.doi.org/10.3847/1538-4357/abf7ba.

M. Voetberg, Ashia Livaudais, Becky Nevin, Omari Paul, and Brian Nord. Deepbench: A simulation package for physical benchmarking data. *JOSS*, 6 2023. doi: 10.2172/1989920. URL https://www.osti.gov/biblio/1989920.

Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L Masters, Vihang Mehta, Brooke D Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 09 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab2093.

Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M M Scaife, Chris Lintott, Karen L Masters, James M Dawson, Hugh Dickinson, Lucy Fortson, Izzy L Garland, Kameswara Mantha, David O'Ryan, Jürgen Popp, Brooke Simmons, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DESI: Detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys. *Monthly Notices of the Royal Astronomical Society*, 526(3):4768–4786, 09 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad2919.

Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J. Lintott, Kameswara Mantha, Devina Mohan, David O'Ryan, and Inigo V. Slijepevic. Scaling laws for galaxy images, 2024. URL https://arxiv.org/abs/2404.02973.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art, 2024. URL https://arxiv.org/abs/2308.01222.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, October 2018. ISSN 0925-2312. doi: 10.1016/j.neucom.2018.05.083. URL http://dx.doi.org/10.1016/j.neucom.2018.05.083.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics, 2022a. URL https://arxiv.org/abs/2201.11969.

Rui Wang, Robin Walters, and Rose Yu. Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting, 2022b. URL https://arxiv.org/abs/2206.09450.

Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1911.08251.

Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1458.

Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), July 2020. ISSN 2157-6904. doi: 10.1145/3400066. URL https://doi.org/10.1145/3400066.

Appendix A. Model Architectures

Layers	**Properties**	**Stride**	**Padding**	**Output Shape**
Input	3 x 100 x 100			
Conv2D (w/ BatchNorm2D)	Filters: 8 Kernel: 5x5 Activation: ReLU	1	2	(8, 100, 100)
MaxPool2D	Kernel: 2x2	2	0	(8, 50, 50)
Dropout	p=0.2			(8, 50, 50)
Conv2D (w/ BatchNorm2D)	Filters: 16 Kernel: 3x3 Activation: ReLU	1	1	(16, 50, 50)
MaxPool2D	Kernel: 2x2	2	0	(16, 25, 25)
Dropout	p=0.2			(16, 25, 25)
Conv2D (w/ BatchNorm2D)	Filters: 32 Kernel: 3x3 Activation: ReLU	1	1	(32, 25, 25)
MaxPool2D	Kernel: 2x2	2	0	(32, 12, 12)
Dropout	p=0.2			(32, 12, 12)
Linear (w/ LayerNorm)	Input Dimension: 4608 Output Dimension: 256 Activation: None			(256)
Linear	Input Dimension: 256 Output Dimension: 6 Activation: None			(6)

Table A1. CNN architecture used with the GZ Evo dataset. For other experiments, the architecture only differs in the input dimension, dimension matching after convolutional layers, and output dimension (logits) size.

Layers	**Properties**	**Stride**	**Padding**	**Output Shape**
Input	3 x 100 x 100			
R2Conv (w/ InnerBatchNorm)	Filters: 64 Kernel: 5x5 Activation: ReLU	1	2	(64, 100, 100)
PointwiseMaxPool2D	Kernel: 2x2	2	0	(64, 50, 50)
PointwiseDropout	p=0.2			(64, 50, 50)
R2Conv (w/ InnerBatchNorm)	Filters: 128 Kernel: 3x3 Activation: ReLU	1	1	(128, 50, 50)
PointwiseMaxPool2D	Kernel: 2x2	2	0	(128, 25, 25)
PointwiseDropout	p=0.2			(128, 25, 25)
R2Conv (w/ InnerBatchNorm)	Filters: 256 Kernel: 3x3 Activation: ReLU	1	1	(256, 25, 25)
PointwiseMaxPool2D	Kernel: 2x2	2	0	(256, 12, 12)
PointwiseDropout	p=0.2			(256, 12, 12)
GroupPooling				(32, 12, 12)
Linear (w/ LayerNorm)	Input Dimension: 4608 Output Dimension: 256 Activation: None			(256)
Linear	Input Dimension: 256 Output Dimension: 6 Activation: None			(6)

Table A2. D_4 ENN architecture used with the GZ Evo dataset. For other experiments, the architecture only differs in the input dimension, dimension matching after convolutional layers, and output dimension (logits) size.