Towards Explaining Distribution Shifts

Sean Kulinski [1] **David I. Inouye** [1]

Abstract

A distribution shift can have fundamental consequences such as signaling a change in the operating environment or significantly reducing the accuracy of downstream models. Thus, understanding distribution shifts is critical for examining and hopefully mitigating the effect of such a shift. Most prior work focuses on merely detecting if a shift has occurred and assumes any detected shift can be understood and handled appropriately by a human operator. We hope to aid in these manual mitigation tasks by explaining the distribution shift using interpretable transportation maps from the original distribution to the shifted one. We derive our interpretable mappings from a relaxation of optimal transport, where the candidate mappings are restricted to a set of interpretable mappings. We then inspect multiple quintessential use-cases of distribution shift in real-world tabular, text, and image datasets to showcase how our explanatory mappings provide a better balance between detail and interpretability than baseline explanations by both visual inspection and our PercentExplained metric.

1. Introduction

Most real-world environments are constantly changing, and in many situations, understanding how a specific operating environment has changed is crucial to making decisions respective to such a change. Such a change might be due to a new data distribution seen in deployment which causes a machine-learning model to begin to fail. Another example is a decrease in monthly sales data which could be due to a temporary supply chain issue in distributing a product or could mark a shift in consumer preferences. When these changes are encountered, the burden is often placed on a human operator to investigate the shift and determine the appropriate reaction, if any, that needs to be taken. In this work, our goal is to aid these operators in providing an explanation of such a change.

This ubiquitous phenomenon of having a difference between related distributions is known as distribution shift. Much prior work focuses on *detecting* distribution shifts; however, there is little prior work that looks into *understanding* a detected distribution shift. As it is usually solely up to an operator investigating a flagged distribution shift to decide what to do next, understanding the shift is critical for the operator to more efficiently mitigate any harmful effects of the distribution shift. Due to the fact that there are no cohesive methods for understanding distribution shifts, as well as, the potential high complexity of distribution shifts (e.g., (Koh et al., 2021)), this important manual investigation task can be daunting. The current de facto standard in analyzing a shift in tabular data is to look at how the mean of the original, *source*, distribution shifted to the new, *target*, distribution. However, this simple explanation can miss crucial shift information due to being a coarse summary (e.g., Figure 2) or, in high-dimensional regimes, can be uninterpretable. Thus, there is a need for methods that can automatically provide detailed, yet interpretable, information about a detected shift which ultimately can lead to actionable insights about the shift.

Therefore, we propose a novel framework for explaining distribution shifts, such as showing how features have changed or how groups within the distributions have shifted. Since a distribution shift can be seen as a movement from a source distribution $\boldsymbol{x} \sim P_{src}$ to a target distribution $\boldsymbol{y} \sim P_{tgt}$, we define a distribution shift explanation as a transport map $T(\boldsymbol{x})$ which maps a point in our source distribution onto a point in our target distribution. For example, under this framework, the typical distribution shift explanation via mean shift can be written as $T(\boldsymbol{x}) = \boldsymbol{x} + (\mu_{\boldsymbol{y}} - \mu_{\boldsymbol{x}})$. Intuitively, these transport maps can be thought of as a functional approximation of how the source distribution could have moved in a distribution space to become our target distribution. However, without making assumptions on the type of shift, there exist many possible mappings that explain the shift (see subsection A.1 for examples). Thus, we leverage prior optimal transport work to define an ideal distribution shift explanation and develop practical algorithms for finding and validating such maps.

[1]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. Correspondence to: David I. Inouye <dinouye@purdue.edu>, Sean Kulinski <skulinsk@purdue.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

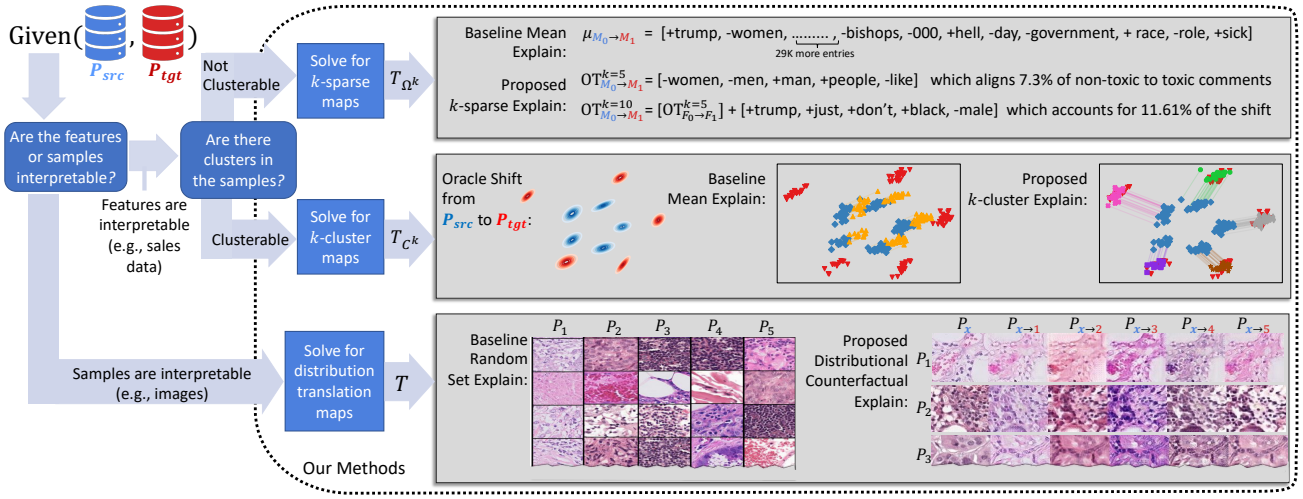


Figure 1: An overall look at our approach to explaining distribution shifts, where given a source P_{src} and shifted P_{tgt} dataset pair, a user can choose to explain the distribution shift using k-sparse maps (which are best suited for high dimensional or feature-wise complex data), k-cluster maps (for tracking how heterogeneous groups change across the shift), or distribution translation maps (for data which has uninterpretable raw features such as images). For details on the results seen in the three boxes, please see experiments in section 5 and Appendix C.

We summarize our contributions as follows:

- In section 3, we define intrinsically interpretable transport maps by constraining a relaxed form of the optimal transport problem to only search over a set of interpretable mappings and suggest possible interpretable sets. Also, we suggest methods for explaining image-based shifts such as distributional counterfactual examples.
- In section 4, we develop practical methods and a Percent-Explained metric for finding intrinsically interpretable mappings which allow us to adjust the interpretability of an explanation to fit the needs of a situation.
- In section 5, we show empirical results on real-world tabular, text, and image-based datasets demonstrating how our explanations can aid an operator in understanding how a distribution has shifted.

2. Related Works

The characterization of the problem of distribution shift has been extensively studied (Quiñonero-Candela et al., 2009; Storkey, 2009; Moreno-Torres et al., 2012) via breaking down a joint distribution $P(\boldsymbol{x}, y)$ of features $\boldsymbol{x}$ and outputs y, into conditional factorizations such as $P(y|\boldsymbol{x})P(\boldsymbol{x})$ or $P(\boldsymbol{x}|y)P(y)$. For covariate shift (Sugiyama et al., 2007) the $P(\boldsymbol{x})$ marginal differs from source to target, but the output conditional $P(y|\boldsymbol{x})$ the same, while label shift (also known as prior probability shift) (Zhang et al., 2013; Lipton et al., 2018) is when the $P(y)$ marginals differ from source to target, but the full-feature conditional $P(\boldsymbol{x}|y)$ remains the same. In this work, we refer to general problem distribution shift, i.e. a shift in the joint distribution (with no distinction between y and $\boldsymbol{x}$), and leave applications of explaining specific sub-genres of distribution shift to future work.

As far as we are aware, this is the first work specifically tackling explaining distribution shifts, thus there are no accepted methods, baselines, or metrics for distribution shift explanations. However, there are distinct works that can be applied to explain distribution shifts. For example, one could use feature attribution methods (Saarela & Jauhiainen, 2021; Molnar, 2020) on a domain/distribution classifier (e.g., Shanbhag et al. (2021) uses Shapley values (Shapley, 1997) to explain how changing input feature distributions affect a classifier's behavior), or once could find samples which are most illustrative of the differences between distributions (Brockmeier et al., 2021). Additionally, one could use counterfactual generation methods (Karras et al., 2019; Sauer & Geiger, 2021; Pawelczyk et al., 2020) and apply them for "distributional counterfactuals" which would show what a sample from P_{tgt} would have looked like if it instead came from P_{src} (e.g., Pawelczyk et al. (2020) uses a classifier-guided VAE to generate class counterfactuals on tabular data). We explore this distributional counterfactual explanation approach in subsection 3.4.

A sister field is that of detecting distribution shifts. This is commonly done using methods such as statistical hypothesis testing of the input features (Nelson, 2003; Rabanser et al., 2018; Quiñonero-Candela et al., 2009), training a domain classifier to test between source and non-source domain samples (Lipton et al., 2018), etc. In Kulinski et al. (2020); Budhathoki et al. (2021), the authors attempt to provide more information beyond the binary "has a shift occurred?" via localizing a shift to a subset of features or causal mechanisms. Kulinski et al. (2020) does this by introducing the

notion of Feature Shift, which first detects if a shift has occurred and if so, localizes that shift to a specific subset of features that have shifted from source to target. In Budhathoki et al. (2021), the authors take a causal approach via individually factoring the source and target distributions into a product of their causal mechanisms (i.e. a variable conditioned on its parents) using a shared causal graph, which is assumed to be known *a priori*. Then, the authors "replace" a subset of causal mechanisms from P_{src} with P_{tgt}, and measure divergence from P_{src} (i.e. measuring how much the subset change affects the source distribution). Both of these methods are still focused on detecting distribution shifts (via identifying shifted causal mechanisms or feature-level shifts), unlike explanatory mappings which help explain *how* the data has shifted.

3. Explaining Shifts via Transport Maps

The underlying assumption of distribution shift is that there exists a relationship between the source and target distributions. From a distributional standpoint, we can view distribution shift as a *movement*, or transportation, of samples from the source distribution P_{src} to the target distribution P_{tgt}. Thus, we can capture this relationship between the distributions via a transport map T from the source distribution to the target, i.e., if $\boldsymbol{x} \sim P_{src}$, then $T(\boldsymbol{x}) \sim P_{tgt}$. If this mapping is understandable to an operator investigating a distribution shift, this can serve as an explanation to the operator on what changed between the environments; thus allowing for more effective reactions to the shift. Therefore, in this work, we define a distribution shift explanation as: *finding an interpretable transport map T which approximately maps a source distribution P_{src} onto a target distribution P_{tgt} such that $T_\sharp P_{src} \approx P_{tgt}$*. Similar to ML model interpretability (Molnar, 2020), an interpretable map can either be one that is intrinsically interpretable (subsection 3.1) or a mapping that is explained via post-hoc methods such as sets of input-output pairs (subsection 3.4).

3.1. Intrinsically Interpretable Transportation Maps

To find such a mapping between distributions, it is natural to look to Optimal Transport (OT) and its extensive prior work in this field (Cuturi, 2013; Arjovsky et al., 2017; Torres et al., 2021; Peyré & Cuturi, 2019). An OT mapping given a transportation cost function c is a method of optimally moving points from one distribution to align with another distribution and is defined as:

$$T_{OT} \coloneqq \arg\min_{T} \mathbb{E}_{P_{src}}\left[c(\boldsymbol{x}, T(\boldsymbol{x}))\right] \text{ s.t. } T_\sharp P_{src} = P_{tgt}$$

where $T_\sharp P_{src}$ is the pushforward operator that can be viewed as applying T to all points in P_{src}, and $T_\sharp P_{src} = P_{tgt}$ is the marginal constraint, which means the pushforward distribution must match the target distribution. OT is a natural starting point for shift explanations as it allows for a rich geometric structure on the space of distributions, and by finding a mapping that minimizes a transport cost we are forcing our mapping to retain as much information about the original $\boldsymbol{x}$ samples when aligning P_{src} and P_{tgt}. For more details about OT, please see (Villani, 2009; Peyré & Cuturi, 2019).

However, since OT considers all possible mappings which satisfy the marginal constraint, this means the resulting T_{OT} can be arbitrarily complex and thus possibly uninterpretable as a shift explanation. We can alleviate this by restricting the candidate transport maps to belong to a set of user-defined interpretable mappings Ω. However, this problem can be infeasible if Ω does not contain a mapping that satisfies the marginal alignment constraint. Therefore, we can use Lagrangian relaxation to relax the marginal constraint, giving us an *Interpretable Transport* mapping T_{IT}:

$$T_{IT} \coloneqq \arg\min_{T \in \Omega} \mathbb{E}_{P_{src}}\left[c(\boldsymbol{x}, T(\boldsymbol{x}))\right] + \lambda\, \phi(P_{T(\boldsymbol{x})}, P_{tgt}) \tag{1}$$

where $\phi(\cdot,\cdot)$ is a distribution divergence function (e.g., KL or Wasserstein). In this paper, we will assume c is the squared Euclidean cost and $\phi(\cdot,\cdot)$ is the squared Wasserstein-2 metric, unless stated otherwise. Due to the heavily complex and context-specific nature of distribution shift, it is likely that Ω would be initialized based on context. However, we suggest two *general* methods in the next section as a starting point and hope that future work can build upon this framework for specific contexts.

3.2. Intrinsically Interpretable Transport Sets

The current common practice for explaining distribution shifts is comparing the means of the source and the target distributions. The mean shift explanation can be generalized as $\Omega_{\text{vector}} = \{T : T(\boldsymbol{x}) = \boldsymbol{x} + \delta\}$ where δ is a constant vector and mean shift being the specific case where δ is the difference of the source and target means. By letting δ be a function of $\boldsymbol{x}$, which further generalizes the notion of mean shift by allowing each point to move a variable amount per dimension, we arrive at a transport set that includes any possible mapping $T : \mathbb{R}^D \to \mathbb{R}^D$. However, even a simple transport set like Ω_{vector} can yield uninterpretable mappings in high dimensional regimes (e.g., a shift vector of over 100 dimensions). To combat this, we can constrain the complexity of a mapping by forcing it to only move points along a specified number of dimensions, which we call *k-Sparse Transport*.

k-Sparse Transport: For a given class of transport maps, Ω and a given $k \in \{1, ..., D\}$, we can find a subset $\Omega^{(k)}_{sparse}$ which is the set of transport maps from Ω which only transport points along k dimensions or less. Formally, we define an active set $\mathcal{A}$ to be the set of dimensions along which a given T moves points: $\mathcal{A}(T) \triangleq \{j \in \{1, \ldots, D\} : \exists \boldsymbol{x}, T(\boldsymbol{x})_j - x_j \neq 0\}$. Then, we define $\Omega^{(k)}_{sparse} = \{T \in \Omega : |\mathcal{A}(T)| \leq k\}$.

k-sparse transport is most useful in situations where a distribution shift has happened along a subset of dimensions, such as explaining a shift where some sensors in a network are picking up a change in an environment. However, in cases where points shift in different directions based on their original value, e.g. when investigating how a heterogeneous population responded to an advertising campaign, k-sparse transport is not ideal. Thus, we provide a shift explanation that breaks the source and target distributions into k sub-populations and provides a vector-based shift explanation per sub-population, which we call *k-Cluster Transport*.

k-Cluster Transport: Given a $k \in \{1, \ldots, D\}$ we define k-cluster transport to be a mapping which moves each point $\boldsymbol{x}$ by constant vector which is specific to $\boldsymbol{x}$'s cluster. More formally, we define a labeling function $\sigma(\boldsymbol{x}; M) \triangleq \arg\min_j \|\boldsymbol{m}_j - \boldsymbol{x}\|_2$, which returns the index of the column in M (i.e. the label of the cluster) which $\boldsymbol{x}$ is closest to. With this, we define $\Omega^{(k)}_{\text{cluster}} = \left\{T : T(\boldsymbol{x}) = \boldsymbol{x} + \delta_{\sigma(\boldsymbol{x};M)}, M \in \mathbb{R}^{D\times k}, \Delta \in \mathbb{R}^{D\times k}\right\}$, where δ_j is the jth column of Δ.

Since measuring the exact interpretability of a mapping is heavily context-dependent, we can instead use k in the above transport maps to define a partial ordering of interpretability of mappings *within* a class of transport maps. Let k_1 and k_2 be the size of the active sets for k-sparse maps (or the number of clusters for k-cluster maps) of T_1 and T_2 respectively. If $k_1 \leq k_2$, then Inter$(T_1) \geq$ Inter(T_2), where Inter(T) is the interpretability of shift explanation T. For example, we claim the interpretability of a $T_1 \in \Omega^{(k=10)}_{sparse}$ is greater than (or possibly equal to) the interpretability of a $T_2 \in \Omega^{(k=100)}_{sparse}$ since a shift explanation in Ω which moves points along only 10 dimensions is more interpretable than a similar mapping which moves points along 100 dimensions. A similar result can be shown for k-cluster transport since an explanation of how 5 clusters moved under a shift is less complicated than an explanation of how 10 clusters moved. The above method allows us to define a partial ordering on interpretability without having to determine the absolute value of interpretability of an individual explanation T, as this requires expensive context-specific human evaluations, which are out of scope for this paper.

3.3. Intrinsically Interpretable Maps For Images

To find interpretable transport mappings for images, we could first project P_{src} and P_{tgt} onto a low-dimensional *interpretable* latent space (e.g., a space which has disentangled and semantically meaningful dimensions) and then apply the methods above in this *latent* space. Concretely, let us denote the (pseudo-)invertible encoder as $g : \mathbb{R}^D \to \mathbb{R}^{D'}$ where $D' < D$ (e.g., an autoencoder). Given this encoder, we define our set of high dimensional interpretable transport maps: $\Omega_{\text{high-dim}} := \left\{T : T = g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right), \tilde{T} \in \Omega, g \in \mathcal{I}\right\}$ where Ω the set of interpretable mappings (e.g., k-sparse mappings) and $\mathcal{I}$ is the set of (pseudo-)invertible functions with an interpretable (i.e. semantically meaningful) latent space. Finally, given an interpretable $g \in \mathcal{I}$, this gives us *High-dimensional Interpretable Transport*: T_{HIT}.

As seen in the Stanford Wilds dataset (Koh et al., 2021), which contains benchmark examples of real-world image-based distribution shifts, image-based shifts can be immensely complex. In order to provide an adequate intrinsically interpretable mapping explanation of a distribution shift in high dimensional data (e.g., images), multiple new advancements must first be met (e.g., finding a disentangled latent space with semantically meaningful dimensions, approximating high dimensional empirical optimal transport maps, etc.), which are out of scope of this paper. We further explore details about T_{HIT}, its variants, and the results of using T_{HIT} to explain Colorized-MNIST in Appendix D, and we hope future work could build upon this framework.

3.4. Post-Hoc Explanations of Image-Based Mappings via Counterfactual Examples

As mentioned above, in some cases, solving for an interpretable latent space can be too difficult or costly, and thus a shift cannot be expressed by an interpretable mapping function. However, if the samples themselves are easy to interpret (e.g., images), we can still explain a transport mapping by visualizing translated samples. Specifically, we can remove the interpretable constraint on the mapping itself and leverage methods from the unpaired Image-to-Image translation (I2I) literature to translate between the source and target domain while preserving the content. For a comprehensive summary of the recent I2I works and methods, please see (Pang et al., 2021).

Once an I2I mapping is found, to serve as an explanation, we can provide an operator with a set of counterfactual pairs $\{(\boldsymbol{x}, T(\boldsymbol{x})) : \boldsymbol{x} \sim P_{src}, T(\boldsymbol{x}) \sim P_{tgt}\}$. Then, by determining what commonly stays invariant and what commonly changes across the set of counterfactual pairs, this can serve as an explanation of how the source distribution shifted to the target distribution. While more broadly applicable, this approach could put a higher load on the operator than an intrinsically interpretable mapping approach.

4. Practical Methods for Finding and Validating Shift Explanations

In this section, we discuss practical methods for finding these maps via empirical OT (Sec. 4.1, 4.2, and 4.3) and introduce a PercentExplained metric which can assist the operator in selecting the hyperparameter k in k-sparse and k-cluster transport (Sec. 4.4).

4.1. Empirical Interpretable Transport Upper Bound

As the divergence term in our interpretable transport objective (Equation 1) can be computationally-expensive to optimize in practice, we propose to optimize the following simplification, which simply computes the difference between the map and the sample-based OT solution T_{OT} (which can be computed efficiently for samples or approximated via the Sinkhorn algorithm (Cuturi, 2013)):

$$\arg\min_{T\in\Omega} \frac{1}{N}\sum_{i=1}^{N} c\big(\boldsymbol{x}^{(i)}, T(\boldsymbol{x}^{(i)})\big) + \lambda d\big(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)})\big) \tag{2}$$

where d is the squared ℓ_2 function. Notably, the divergence value in Equation 1 is replaced with the average over a sample-specific distance between $T(\boldsymbol{x})$ and the optimal transport mapping $T_{OT}(\boldsymbol{x})$. This is computationally attractive as the optimal transport solution only needs to be calculated once, rather than calculating the Wasserstein distance once per iteration as would be required if directly optimizing the Interpretable Transport problem. Additionally, we prove in subsection A.2 that the second term in Equation 2 is an upper bound when the divergence is the squared Wasserstein distance, i.e., when $\phi = W_2^2$.

4.2. Finding k-Sparse Maps

The k-sparse algorithm can be broken down into two steps. First, given k, we estimate the active set $\mathcal{A}$ by simply taking the k dimensions with the largest difference of means between two distributions. This is a simple approach that avoids optimization over an exponential number of possible subsets for $\mathcal{A}$ and can be optimal for some cases, as explained below. Second, given the active set $\mathcal{A}$, we need to estimate the map. While estimating k-sparse solutions to the original interpretable transport problem (Equation 1) is challenging, we prove that the solution with optimal alignment to the upper bound above (Equation 2) can be computed in closed-form for two special cases. If the optimization set is restricted to only shifting the mean, i.e., $\Omega^{(k)} = \Omega^{(k)}_{vector}$, then the solution with optimal alignment is:

$$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} x_j + (\mu_j^{\text{tgt}} - \mu_j^{\text{src}}), & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases}, \tag{3}$$

where μ^{src} and μ^{tgt} are the mean of the source and target distributions respectively. Similarly, if $\Omega^{(k)}$ is unconstrained except for sparsity, then the solution with optimal alignment is simply:

$$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} [T_{OT}(\boldsymbol{x})]_j, & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases}, \tag{4}$$

where $[T_{OT}(\boldsymbol{x})]_j$ is the j-th coordinate of the sample-based OT solution. The proofs of alignment optimality w.r.t. the divergence upper bound in Equation 2 are based on decomposability of the squared Euclidean distance and can be found in Appendix A. The final algorithm for both sparse maps can be found in Algorithm 1.

Algorithm 1 Finding k-Sparse Maps

Input: Domain datasets $X \in \mathbb{R}^{N\times D}$ and $Y^{N\times D}$ with N samples of dimensionality D each, the desired sparsity k, and interpretable set type, i.e., Ω.

// Select active set based on means
$\mu^{\text{diff}} \leftarrow \mu^{\text{tgt}} - \mu^{\text{src}} = \frac{1}{N}\sum_{i=1}^{N} Y_i - \frac{1}{N}\sum_{i=1}^{N} X_i$
$\mathcal{A} \leftarrow \text{TopKIndices}(\text{abs}(\mu^{\text{diff}}), k)$
// Create dimension-wise maps based on active set
if $\Omega = \Omega_{vector}$ **then**
$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} x_j + \mu_j^{\text{diff}}, & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases}$
else
$T_{OT}(\cdot) \leftarrow \text{OptimalTransportAlg}(X, Y)$
$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} [T_{OT}(\boldsymbol{x})]_j, & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases}$
end if
Output: $T(\cdot)$

4.3. Finding k-Cluster Maps

Similar to k-sparse maps, we split this algorithm into two parts: (1) estimate pairs of source and target clusters and then (2) compute mean shift for each pair of clusters. For the first step, naïvely one might expect that independent clustering on each domain dataset followed by post-hoc pairing of these clusters may be sufficient. However, this could yield very poor clustering pairs that are significantly mismatched because the domain-specific clustering may not be optimal in terms of the alignment objective. For example, the source domain may have one large and one small cluster and the target domain could have equal-sized clusters. Therefore, it is important to cluster the source and domain samples jointly. To estimate paired (i.e., dependent) clusterings of the source and target domain samples, we first find the OT mapping from source to target. We then cluster an paired dataset formed by concatenating each source sample with its OT mapped sample (which actually corresponds to one of the target samples). The clustering on these paired samples gives paired cluster centroids for the source and target, denoted μ_ℓ^{src} and μ_ℓ^{tgt} respectively, which we use to construct a cluster-specific mean shift map defined as:

$$T(\boldsymbol{x}) = \boldsymbol{x} + (\mu^{\text{tgt}}_{\sigma(\boldsymbol{x})} - \mu^{\text{src}}_{\sigma(\boldsymbol{x})}) \tag{5}$$

where $\sigma(\boldsymbol{x}) = \arg\min_\ell \|\boldsymbol{x} - \mu_\ell^{\text{src}}\|_2^2$ is the cluster label function. This map applies a simple shift to every source domain cluster to map to the target domain. Algorithm 2 shows pseudo-code for both steps in our k-cluster method.

Algorithm 2 Solving for k-Cluster Mappings

Input: Domain datasets $X \in \mathbb{R}^{N \times D}$ and $Y^{N \times D}$ with N samples of dimensionality D each and the desired number of clusters k.

// Compute sample-based optimal transport map
$T_{OT}(\cdot) \leftarrow \text{OptimalTransportAlg}(X, Y)$
// Compute paired clustering
$Z \leftarrow [X, T_{OT}(X)] \in \mathbb{R}^{N \times 2D}$
$[\mu_1, \cdots, \mu_k]^T \leftarrow \text{KMeansClust}(Z, k) \in \mathbb{R}^{k \times 2D}$
// Extract paired source and target centroids
$\forall \ell \in \{1, \cdots, k\}, \mu_\ell^{\text{src}} = [\mu_{\ell,1}, \cdots, \mu_{\ell,D}]^T \in \mathbb{R}^D$
$\forall \ell \in \{1, \cdots, k\}, \mu_\ell^{\text{tgt}} = [\mu_{\ell,D+1}, \cdots, \mu_{\ell,2D}]^T \in \mathbb{R}^D$
// Setup final cluster-based map
$\sigma(\boldsymbol{x}) = \arg\min_\ell \|\boldsymbol{x} - \mu_\ell^{\text{src}}\|_2^2$ *// Clust. label func*
$T(\boldsymbol{x}) = \boldsymbol{x} + (\mu_{\sigma(\boldsymbol{x})}^{\text{tgt}} - \mu_{\sigma(\boldsymbol{x})}^{\text{src}})$
Output: $T(\cdot)$

4.4. Interpretability as a Hyperparameter

We now discuss how the k hyperparameter in k-sparse and k-cluster maps can be adjusted to allow a user to automatically change the level of interpretability of a shift explanation as desired. While an optimal shift explanation could be achieved by solving Equation 1, directly defining the set Ω, which should contain both interpretable yet sufficiently expressive maps, can be a difficult task. Thus, we can instead set Ω to be a super-class, such as Ω_{vector} given in subsection 3.2 and adjust k until a $\Omega^{(k)}$ is found which matches the needs of the situation. This allows a human operator to request a mapping with better alignment by increasing k, which correspondingly will decrease the mapping's interpretability, or request a more interpretable mapping by decreasing the complexity (i.e. decreasing k) which will decrease the alignment.

To assist an operator in determining if the interpretability hyperparameter should be adjusted, we introduce a *Percent-Explained* metric, which we define to be:

$$\text{PE}(P_{src}, P_{tgt}, T) := \frac{W_2^2(P_{src}, P_{tgt}) - W_2^2(T_\sharp P_{src}, P_{tgt})}{W_2^2(P_{src}, P_{tgt})} \tag{6}$$

where $W_2^2(\cdot, \cdot)$ is the squared Wasserstein-2 distance between two distributions and PE is shorthand for PercentExplained. By rearranging terms we get $1 - \frac{W_2^2(T_\sharp P_{src}, P_{tgt})}{W_2^2(P_{src}, P_{tgt})}$, which shows this metric's correspondence to the statistics coefficient of determination R^2, where $W_2^2(T_\sharp P_{src}, P_{tgt})$ is analogous to the residual sum of squares and $W_2^2(P_{src}, P_{tgt})$ is similar to the total sum of squares. This gives an approximation of how much a current shift explanation T accurately maps onto a target distribution. This can be seen as a normalization of a mapping's fidelity with the extremes being $T_\sharp P_{src} = P_{tgt}$, which fully captures a shift, and $T = \text{Id}$, which does not move the points at all. When provided this metric along with a shift explanation, an operator can decide whether to accept the explanation (e.g., the PercentExplained is sufficient and T is still interpretable) or reject the explanation and adjust k.

5. Experiments

In this section, we study the performance of our methods when applied to real-world data.[1] For gaining intuition on different explanation techniques, we point the reader to Appendix C where we present experiments on simple simulated shifts. We first present results using k-cluster transport to explain the difference between different groups of the male population and groups of the female population in the U.S. Census "Adult Income" dataset (Kohavi & Becker, 1996). We then use k-sparse transport to explain shifts between toxic and non-toxic comments across splits from the Stanford WILDS distribution shift benchmark (Koh et al., 2021) version of the "CivilComments" Dataset (Borkan et al., 2019). Finally, we use distributional counterfactuals to explain the high-dimensional shifts between histopathology images from different hospitals as seen in the WILDS Camelyon17 dataset (Bandi et al., 2018).

Adult Income Dataset This dataset originally comes from the United States 1994 Census database and is commonly used to predict whether a person's annual income exceeds \$50k using 14 demographic features. Similar to (Budhathoki et al., 2021), we consider a subset of non-redundant features: age, years of education (where 12+ is beyond high school), and income (which is encoded as 1 if the person's annual income is greater than \$50k and 0 if it is below). We then split this dataset along the sex dimension, and define our source distribution as the male population and the target as the female population. In order to find the set of paired clusters, we first standardize a copy of the data to have zero mean and unit variance across all features, where the μ and σ used for the standardization are found via the feature-wise mean and standard deviation of the source distribution and perform clustering in the standardized joint space using the method described in Section 4.3. The k clustering labels are then used to label points to clusters in the original (unstandardized) data space.

Suppose our role is a researcher seeking to implement a social program targeting gender inequalities. We could compare the means of the male/female distributions, which shows on average a 20% lower chance of having an annual income above \$50k when moving from the male population to the female population. Additionally, we could train a classifier to predict between male/female data points and use a feature importance measurement tool like Shapley values (Lundberg & Lee, 2017) to determine that income is a main differing feature (**insight 1** from Figure 2). How-

[1] Code to recreate the experiments can be found at https://github.com/inouye-lab/explaining-distribution-shifts.

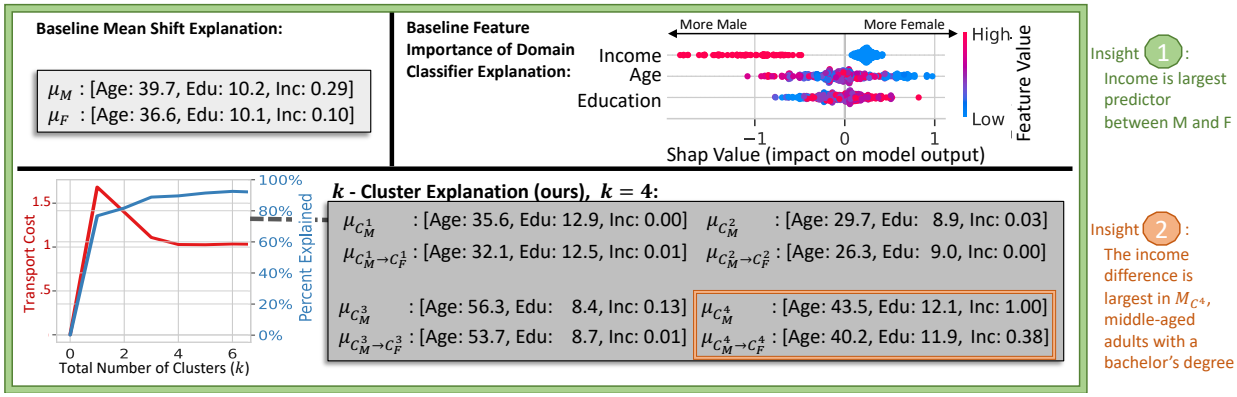


Figure 2: Using k-cluster transport (bottom) to explain the shift from the male population to the female population of the Adult Income dataset allows us to capture how heterogeneous groups within the dataset moved. For example, while all three methods show that income is indeed the largest difference between M and F for this dataset (**insight 1**), only the k-cluster-based explanation reveals **insight 2**, that the income disparity is most prevalent between middle-aged males and females with a bachelor's degree (edu=12) seen in C^4.

ever, suppose we want to dig deeper. We could instead use k-cluster transport to see how heterogeneous subgroups shifted across a range of clusters, as seen in Figure 2. If we accept the explanation at $k = 4$ (since beyond this, the marginal advantage of adding an additional cluster is minimal in terms of both transport cost and PercentExplained), we can now make **insight 2**. Here, $\mu_{C_M^4 \to C_F^4}$ shows the income difference is significantly larger between middle-aged males/females with a bachelor's degree (a decrease from nearly 100% high-income likelihood to only a 38% chance). While **insight 1** validates the need for our social program, **insight 2** (which is hidden in both the mean-shift and distribution classifier explanations) provides a significantly narrower scope for us to focus our efforts in, thus allowing for swifter action.

Civil Comments Dataset Here we present results using k-sparse shifts to explain the difference between three splits of the CivilComments dataset (Borkan et al., 2019) from the WILDS datasets (Koh et al., 2021). This dataset consists of comments scraped from the internet where each comment is paired with a binary toxicity label and demographic information pertaining to the content of the comment. If we were an operator trying to see how the comments and their toxicity change across targeted demographics, we could create three splits: {F, M}, {F_0, F_1}, and {M_0, M_1}, where F represents all female comments, M are all male comments, and F_0, F_1 are nontoxic, toxic female comments, respectively (and likewise for males). We can explain these three splits using vanilla mean shift, a k-sparse mean shift (k-μ), and k-sparse OT (k-OT) shift explanations, as seen in Table 1 which shows results for the unigrams which the maximize the alignment between the unigram distributions created for each split. The baseline vanilla mean-shift explanation yields all 30K features at once (with no guide for truncating), while the k-sparse shifts provide explanations up to k words as well as a corresponding PercentExplained to aid in determining if additional words should be added to the explanation. Note that for k-μ explanations, when transporting a word, that word is added equally to all comments in P_{src}, while since k-OT allows for each comment to be shifted optimally (via conditioning on the other words in each comment), thus k-OT can explain significantly more of the shift by transporting the same word (which can highlight words that have complex dependencies such as "don't").

It is clear that performing a vanilla mean shift explanation on the unigram data between splits is unwise due to the high dimensionality of the data and it is unclear when to truncate such an explanation. However, in our approach, by iteratively reporting the shifted unigrams along with the cumulative PercentExplained, a practitioner can better understand the impact each additional word has on the shift explanation. For example, it makes sense that adding "man", "men", and subtracting "woman" were the three unigrams that best aligned the female and male comment distributions and could account for as much as 10% of the shift.

With the k-sparse explanation, **insight 1** suggests that a content moderator is more likely to encounter toxic comments that target individuals rather than groups, and thus a moderator could train a classifier to predict if the object of a comment is an individual or group of people. This quality, which is not obvious at first glance (and especially not from the simple mean shift explanation), may enable more explainable moderation as an explanation about why a comment was removed could state that a comment is targeting an individual as one feature. Additionally, **insight 2** shows that if the moderator's goal is to be equitable across groups, they may want to target words that have roughly equal impact across groups like "stupid". If on the other hand, the moderator's goal is merely to detect any toxicity

Table 1: A baseline vanilla mean shift explanation, k-sparse mean shift explanation, (k-μ-Ex), and k-sparse OT explanations (k-OT-Ex) for the three splits from CivilComments (to save space the baseline is only used for F→M). Each cell represents adding/subtracting a unigram from P_{src} to align it with the comment distribution of P_{tgt} and the respective PercentExplained (excluding the baseline method). For example, in k-μ-Ex(F_0→F_1), adding "stupid" aligns the non-toxic female comments to the toxic female comments and cumulatively explains 0.2% of the shift.

k	Baseline: $\mu(F, M)$
1	+ man
2	+ men
3	- woman
4	+ white
... [29K more entries] ...	
29,553	+ martina
29,554	- diqlmjawsae
29,555	- да
29,556	- bodybuilder
29,557	+ philhiblers

k	k-μ-Ex(F, M)		k-OT-Ex(F, M)	
1	+ man	2.4%	+ man	6.9%
2	+ men	2.4%	+ men	9.2%
3	- woman	2.6%	- woman	10.8%
4	+ white	2.7%	+ white	12.0%
5	+ male	2.8%	- people	13.0%
6	+ black	2.8%	- like	13.9%
7	+ god	2.8%	- just	14.7%
8	- female	2.8%	+ male	15.4%
9	- abortion	2.8%	- don't	16.3%
10	+ males	2.8%	+ god	16.8%

k	k-μ-Ex(F_0, F_1)		k-OT-Ex(F_0, F_1)	
1	+ white	0.1%	+ trump	2.6%
2	- women	0.1%	+ people	3.7%
3	+ like	0.2%	+ woman	4.7%
4	+ stupid	0.2%	+ like	5.7%
5	- church	0.2%	- men	6.6%
6	+ hillary	0.2%	+ just	7.5%
7	+ black	0.2%	+ don't	8.3%
8	+ sex	0.2%	+ white	9.2%
9	- female	0.2%	+ man	9.9%
10	- abortion	0.3%	- think	10.5%

k	k-μ-Ex(M_0, M_1)		k-OT-Ex(M_0, M_1)	
1	+ trump	0.2%	- women	2.8%
2	- women	0.3%	- men	4.1%
3	+ black	0.3%	+ man	5.3%
4	- church	0.3%	+ people	6.3%
5	+ stupid	0.3%	+ like	7.3%
6	+ gay	0.4%	+ trump	8.3%
7	+ racist	0.4%	+ just	9.2%
8	- god	0.4%	+ don't	10.9%
9	- jesus	0.4%	+ black	11.6%
10	+ man	0.4%	- male	12.3%

Insight 1: The decrease (increase) in plural (singular) words suggests individuals are being targeted in toxic comments

Insight 2: Words such as "stupid" are predictive of toxicity for both males and females, while "racist" is more specific to males

regardless of group, this may signal that they should provide group variables such as "gay" or "racist" (which are more predictive of toxicity towards males) to the model as it is an important signal.

Explaining Shifts in H&E Images Across Hospitals We apply this distribution counterfactual approach to the Camelyon17 dataset (Bandi et al., 2018) which is a real-world distribution shift dataset that consists of whole-slide histopathology images from five different hospitals. We use the Stanford WILDS (Koh et al., 2021) variant of the dataset which converts the whole-slide images into over 400 thousand patches. Since each hospital has varying hematoxylin and eosin (H&E) staining characteristics, this, among other batch effects, leads to heterogeneous image distributions across hospitals, as suggested by **insight 1** in Figure 3.

To generate the counterfactual examples, we treat each hospital as a domain and train a StarGAN model (Choi et al., 2018) to translate between each domain. For training, we followed the original training approach seen in (Choi et al., 2018), with the exception that we perform no center cropping. After training, we can generate distribution counterfactual examples by inputting a source image and the label of the target hospital domain to the model. Counterfactual generation was done for all five hospitals and can be seen on the right-hand side of Figure 3. It can be seen that the distributional counterfactuals lead to **insight 2** —that each hospital has a distinct level of staining that seems to be characteristic across samples from the same hospital. For example, P_1 (hospital 1) consists of mostly light staining, and thus transporting to this domain usually involves lightening of the image. Thus, if a practitioner is building a model which should generalize to slides from a new hospital within a network, **insight 2** confirms that controlling for different levels of staining (e.g., color augmentations) is necessary in the training pipeline. We further explore distinctly content-based counterfactuals (as opposed to style changes such as levels of staining) of an image using the CelebA dataset in subsection C.4.

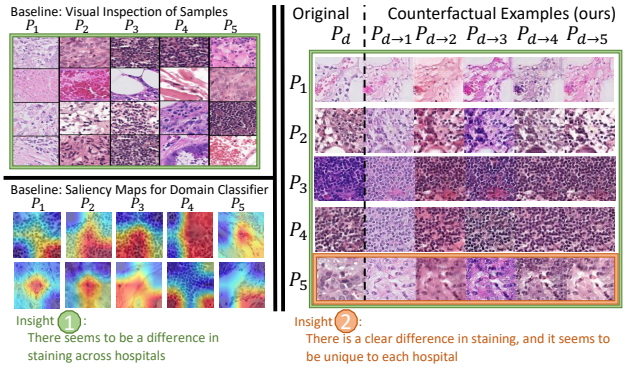


Figure 3: While the baseline method of unpaired samples (top-left) hints that there is a difference in staining (but is relatively unclear), our explanation approach (right) of showing paired counterfactual images translated between the hospital domains (represented as P_1, P_2, ...) quickly leads to the stronger **insight 2** —indeed the staining/coloring differs across the hospital domains *and* the type of staining seems to be consistent and unique for each hospital. For the counterfactual examples, the $(i_{\text{row}}, j_{\text{column}})$ pair represents the pushforward of a sample from domain P_i onto the P_j domain. Using Grad-CAM (Selvaraju et al., 2016) to explain a ResNet-50 (He et al., 2016) domain classifier (bottom-left) does not lead to actionable insights.

6. Discussion and Limitations

Choosing Between k-sparse vs. k-cluster Explanations We first highlight that distribution shift is a highly context-specific problem. Thus, the explanation method will likely depend on the nature of the data (e.g., the data may contain natural subgroups or clusters). If the sparsity or cluster structure is unknown, we suggest following the logical flow for method selection in Figure 1 to determine which shift explanation method to use for their specific context. In short, the k-sparse mappings are useful to allow an operator to see how specific *features* changed from P_{src} to P_{tgt}, and the k-cluster mappings are useful to track how *sub-groups of samples* changed under the distribution shift.

Evaluating Shift Explanations A primary challenge in

developing distribution shift explanations is determining how to evaluate the efficacy of a given explanation in a given context. Evaluating explanations is an active area of research (Robnik-vSikonja & Bohanec, 2018; Molnar, 2020; Doshi-Velez & Kim, 2017) with commonalities such as an explanation should be contrastive, succinct, should highlight abnormalities, and should have high fidelity (Molnar, 2020). For the case of distribution shift explanations, as this is a highly context-dependent problem (dependent on the data setting, task setting, and operator knowledge) and our approach is designed to tackle this problem in general, we do not have a general automated way of measuring whether a given explanation is indeed interpretable. Instead, we provide a general contrastive method that supplies the PercentExplained (approximation of fidelity) and the adjustable k-level of sparse/cluster mappings (which trades off between succinctness and fidelity) but ultimately leaves the task of validating the explanation up to the operator.

Potential Failure Cases Our explanations are meant as a diagnostic tool like most initial data analyses. While our explanations can provide actionable insights, they are one step in the diagnosis process and need to be investigated further before making a decision, especially in high-stakes scenarios. As with most explanation methods, the explanations may be incorrectly interpreted as being *causal* while they likely only show correlations and many hidden confounders could exist. While causal-oriented explanation methods are much more difficult to formulate and optimize, they could provide deeper insights than standard methods. Because OT with squared Euclidean cost is the basis for several of our methods, this could cause a misleading explanation if the scaling of the dimensions does match the operator's expectations (more discussion below). Another example failure case would be using k-cluster transport when no natural clusters exist. In practice, we noticed that StyleGAN can fail to recover large content-based shifts (as opposed to style-based shifts) such as the shift from "wearing hat" $\rightarrow$ "bald" in Celeb-A (as seen in Figure 8). This case is difficult to diagnose, but could be alleviated by using an Image-To-Image translation approach which does not have the style-based biases seen in methods like StyleGAN (Karras et al., 2019).

Motivation for using Optimal Transport and Euclidean cost Because there are many possible mappings between source and target distributions, any useful mapping will need to make some assumptions about the shift. We chose the OT assumption, which can be seen as a simplicity assumption (similar to Occam's Razor) because points are moved as little as possible. Additionally, the OT mapping is unique (even with just samples) and can be computed with fast known algorithms (Peyré & Cuturi, 2019). The squared Euclidean distance is a reasonable cost function when pointwise distances are semantically or causally meaningful (e.g., increasing the education of a person)–as is the case for many shift settings in this work. However, squared Euclidean distance in the raw feature space might not be meaningful for some datasets such as raw pixel values of images. For cases like these, a cost function other than the ℓ_2^2 cost can be used, and we explore this in subsection 3.3 and in detail in Appendix D. In those sections, we look at first applying a semantic-encoding function $g(\mathbf{x})$ which projects $\mathbf{x}$ to a semantically meaningful latent space and then calculate the transportation cost in this meaningful latent space. In general, because OT algorithms can use any distance function, context-dependent cost functions could easily be used within our framework for improved interpretability.

Future Directions For Shift Explanations We believe developing new shift explanation maps and evaluation criteria for specific applications (e.g., explaining the results of biological experiments run with different initial conditions) is a rich area for future work. Also, the PercentExplained metric does not provide information on specifically what is missing from the explanation, i.e., the missing information is a "known unknown". For image-based explanations, the explanation may fail to show certain domain changes that could mislead an operator, i.e., "unknown unknowns". Methods to quantify and analyze these unknowns would improve the trustworthiness of shift explanations. For further discussions of challenges with explaining distribution shifts (e.g., finding an interpretable latent space, approximations of Wasserstein distances in high dimensional regimes, etc.) we point the reader to Appendix B.

7. Conclusion

In this paper, we introduced a framework for explaining distribution shifts using a transport map T between a source and target distribution. We constrained a relaxed form of optimal transport to theoretically define an intrinsically interpretable mapping T_{IT} and introduced two interpretable transport methods: k-sparse and k-cluster transport. We provided practical approaches to calculating a shift explanation, which allows us to use treat interpretability as a hyperparameter that can be adjusted based on a user's need and showed how our methods can help an operator investigate a distribution shift on real-world examples. Both in section 5 and in Appendix C, we show the feasibility of our techniques on many different shift problems to both gain intuition for the different types of shift explanations and to show how our methods can help an operator investigate a distribution shift. We hope our work suggests multiple natural extensions such as using trees as a feature-axis-aligned form of clustering or even other forms of interpretable sets. Given our results and potential ways forward, we ultimately hope our framework lays the groundwork for providing more information to aid in investigations of distribution shift.

Acknowledgements This work was supported in part by ARL (W911NF-2020-221) and ONR (N00014-23-C-1016).

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Bertsimas, D., Orfanoudaki, A., and Wiberg, H. Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138, 2021.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Brockmeier, A. J., Claros-Olivares, C. C., Emigh, M., and Giraldo, L. G. S. Identifying the instances associated with distribution shifts using the max-sliced bures divergence. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Budhathoki, K., Janzing, D., Bloebaum, P., and Ng, H. Why did the distribution change? In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1666–1674. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/budhathoki21a.html.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

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

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Fraiman, R., Ghattas, B., and Svarc, M. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7(2):125–145, 2013.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348. PMLR, 2020.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kohavi, R. and Becker, B. Uci machine learning repository: Adult data set, 1996. URL https://archive.ics.uci.edu/ml/machine-learning-databases/adult.

Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Kulinski, S., Bagchi, S., and Inouye, D. I. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in Neural Information Processing Systems*, 33, 2020.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Mangasarian, O. L. and Wolberg, W. H. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

Molnar, C. *Interpretable machine learning*. Lulu. com, 2020.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Nelson, W. B. *Applied life data analysis*, volume 521. John Wiley & Sons, 2003.

Pang, Y., Lin, J., Qin, T., and Chen, Z. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 2021.

Pawelczyk, M., Broelemann, K., and Kasneci, G. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132, 2020.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607, 2019.

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

Rabanser, S., Günnemann, S., and Lipton, Z. C. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018.

Robnik-vSikonja, M. and Bohanec, M. Perturbation-based explanations of prediction models. In *Human and machine learning*, pp. 159–175. Springer, 2018.

Saarela, M. and Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):1–12, 2021.

Sauer, A. and Geiger, A. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: visual explanations from deep networks via gradient-based localization. 2016. *arXiv preprint arXiv:1610.02391*, 2016.

Shanbhag, A., Ghosh, A., and Rubin, J. Unified shapley framework to explain prediction drift. *arXiv preprint arXiv:2102.07862*, 2021.

Shapley, L. S. A value for n-person games. *Classics in game theory*, 69, 1997.

Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N. D., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5927–5937. Curran Associates, Inc., 2017.

Storkey, A. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Torres, L. C., Pereira, L. M., and Amini, M. H. A survey on optimal transport for machine learning: Theory and applications, 2021.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827. PMLR, 2013.

A. Proofs

A.1. Proof that there are an infinite number of possible mappings between distributions

As stated in the introduction, given two distributions, there exist many possible mappings such that $T_\sharp P_{src} = P_{tgt}$ (it should be noted that here we are speaking of the general mapping problem, not the *optimal* transport problem which can be shown via Brenier's theorem (Peyré & Cuturi, 2019) to have a unique matching for some cases). For instance, given two isometric Gaussian distributions $\boldsymbol{x} \sim \mathcal{N}_1(\mu_1, I)$, $\boldsymbol{y} \sim \mathcal{N}_2(\mu_2, I)$, where I is the Identity matrix, there exist an infinite number of T's such that $T(\boldsymbol{x}) \sim \mathcal{N}_2$. Specifically, any T of the form: $T(\boldsymbol{x}) = \mu_2 + R(\boldsymbol{x} - \mu_1)$, where is R is an arbitrary rotation matrix, will shift $T_\sharp \mathcal{N}_1$ to have a mean of μ_2 and perfectly align the two distributions (since any rotation of an isometric Gaussian will still be an isometric Gaussian).

A.2. Proof that practical interpretable transport objective is an upper bound of theoretic interpretable transport

First, recall our empirical approximation problem for finding T where the second term is an approximation to the Wasserstein distance:

$$\arg\min_{T \in \Omega} \frac{1}{N} \sum_{i=1}^{N} c(\boldsymbol{x}^{(i)}, T(\boldsymbol{x}^{(i)})) + \lambda d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)})) \tag{7}$$

where T_{OT} is the optimal transport solution between our source and target domains with the given c cost function. The empirical average over samples can be viewed as an empirical expectation (which converges to the population expectation as N approaches infinity). Because the Wasserstein distance is well-defined for discrete distributions (like the empirical distribution) and continous distributions, we can simply prove that our approximation is an upper bound for any expectation (empirical or population-level) as follows:

$$\begin{aligned}
W_2^2(P_{T(\boldsymbol{x})}, P_Y) &= \min_{T' : T'_\sharp P_{T(\boldsymbol{x})} = P_{tgt}} \mathbb{E}_{\boldsymbol{z} \sim P_{T(\boldsymbol{x})}} [d(\boldsymbol{z}, T'(\boldsymbol{z}))] && (8)\\
&= \min_{T' : T'_\sharp P_{T(\boldsymbol{x})} = P_{tgt}} \mathbb{E}_{\boldsymbol{x} \sim P_{src}} [d(T(\boldsymbol{x}), T' \circ T(\boldsymbol{x}))] && (9)\\
&\leq \mathbb{E}_{\boldsymbol{x} \sim P_{src}} [d\left(T(\boldsymbol{x}), (T_{OT} \circ T^{-1}) \circ T(\boldsymbol{x})\right)] && (10)\\
&= \mathbb{E}_{\boldsymbol{x} \sim P_{src}} [d(T(\boldsymbol{x}), T_{OT}(\boldsymbol{x}))], && (11)
\end{aligned}$$

where Equation 8 is by definition of Wasserstein distance, Equation 9 is by a change of variables, Equation 10 is by taking $T' = T_{OT} \circ T^{-1}$ (which by construction satisfies the alignment constraint and which must be greater or equal to the minimum), and Equation 11 is merely as simplification. Note that if T is the identity, then the inequality becomes an equality. A similar proof could be used for any W_p^p distance where $p \geq 1$.

A.3. Proof that k-sparse truncated OT minimizes alignment upper bound

In this section, we prove that the best possible alignment objective in terms of the upper bound on Wasserstein distance in Equation 2 is given by the truncated OT solution, i.e., the solution in the limit as $\lambda \to \infty$. This solution is more akin to the constraint-based (i.e., non-Lagragian relaxation) of OT except replacing the alignment metric with the upper bound above. While this enforces the best possible alignment and does not directly consider the transportation cost, the solution is relatively low cost because it based on the OT solution. Additionally, it is the *unique* solution to the problem as $\lambda \to \infty$.

We will now prove that it is the optimal and unique solution. First, let $\boldsymbol{z} = T(\boldsymbol{x})$, $\boldsymbol{z}^{OT} = T_{OT}(\boldsymbol{x})$, and $\boldsymbol{x} \in \mathbb{R}^{N \times D}$. If d is the squared Euclidean distance and we restrict to mappings that only change dimensions in $\mathcal{A}$, then we can decompose the distance term as follows:

$$\sum_{i=1}^{N} d(\boldsymbol{z}_i, \boldsymbol{z}_i^{OT}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{A}} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right)^2 + \underbrace{\sum_{j \notin \mathcal{A}} \left(\boldsymbol{x}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right)^2}_{=\alpha_i \text{ , since constant w.r.t T}} = \sum_{i=1}^{N} \alpha_i + \sum_{j \in \mathcal{A}} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right)^2 . \tag{12}$$

Notice that the sum of squares corresponding to $\mathcal{A}$ is dependent on the mapping while the others are a constant w.r.t. T because T cannot modify any dimensions $j \notin \mathcal{A}$. Given this, we now choose our solution to k-sparse optimal transport as

given in the paper:

$$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} [T_{OT}(\boldsymbol{x})]_j, & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases} \tag{13}$$

where $\mathcal{A}$ is the active set of k dimensions which our k-sparse map T can move points. With this solution, we arrive at the following:

$$\sum_{i=1}^{N} d(\boldsymbol{z}_i, \boldsymbol{z}_i^{OT}) = \sum_{i=1}^{N} \alpha_i + \sum_{j \in \mathcal{A}} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right)^2 = \sum_{i=1}^{N} \alpha_i + \sum_{j \in \mathcal{A}} \sum_{i=1}^{N} \left(\boldsymbol{z}_{i,j}^{OT} - \boldsymbol{z}_{i,j}^{OT}\right)^2 = \sum_{i=1}^{N} \alpha_i \,,$$

where the α_i are positive constants that cannot be reduced by T. Therefore, this is indeed the optimal solution to our empirical interpretable transport problem with the alignment approximation as in Equation 2. This can easily be extended to show that the optimal active set for this case is the one that minimizes $\sum_{i=1}^{N} \alpha_i$, thus the active set should be the k dimensions which have the largest squared difference between $\boldsymbol{x}$ and $\boldsymbol{z}^{OT}$.

To prove uniqueness, we use proof by contradiction. Suppose there exists another optimal solution T' that is distinct from the optimal T given in Equation 13. This would mean that there exists a pair $(\boldsymbol{x}, j)$ such that $[T'(\boldsymbol{x})]_j \neq [T(\boldsymbol{x})]_j = [T_{OT}(\boldsymbol{x})]_j = z_j^{OT}$. However, this would mean that the corresponding term in the summation would be non-zero, i.e., $(z_j - z_j^{OT})^2 > 0$. But this would mean that the overall distance function is greater than the sum of the constants yielding a contradiction to the hypothesis that there could exist another solution. Therefore, the k-sparse solution that optimizes alignment is unique.

A.4. Proof that k-mean shift is the k-vector shift that gives the best alignment

Similar to the previous proof, we consider the solution to Equation 2 when $\lambda \to \infty$, i.e., the optimal alignment solution, but restrict ourselves to the space of vector maps Ω_{vector}. First, we recall the definition of Ω_{vector}:

$$\Omega_{vector}^{(k)} = \{T : T(\boldsymbol{x}) = \boldsymbol{x} + \tilde{\delta}\}, \quad \text{where} \quad \tilde{\delta}_j = \begin{cases} \delta_j, & \text{if } j \in \mathcal{A} \\ 0 & \text{if } j \notin \mathcal{A} \end{cases} , \tag{14}$$

where δ_j for $j \in \mathcal{A}$ are the only learnable parameters. Given this, we decompose the sum of distances similar to the previous proof:

$$\sum_{i=1}^{N} d(\boldsymbol{z}_i, \boldsymbol{z}_i^{OT}) = \sum_{i=1}^{N} \alpha_i + \sum_{j \in \mathcal{A}} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right)^2 \tag{15}$$

$$= \sum_{i=1}^{N} \alpha_i + \sum_{j \in \mathcal{A}} \left((\boldsymbol{x}_{i,j} + \delta_j) - \boldsymbol{z}_{i,j}^{OT}\right)^2 \tag{16}$$

where Equation 15 is from Equation 12 and Equation 16 is by the definition of $\Omega^{(k)}$. Because this is a convex function that decomposes over each coordinate, we can take the derivative and set to zero to solve:

$$\frac{d}{d\delta_j}\left(\sum_{i=1}^{N} d(\boldsymbol{z}_i, \boldsymbol{z}_i^{OT})\right) = \frac{d}{d\delta_j} \sum_{i=1}^{N} \left((\boldsymbol{x}_{i,j} + \delta_j) - \boldsymbol{z}_{i,j}^{OT}\right)^2 = 2 \sum_{i=1}^{N} \left((\boldsymbol{x}_{i,j} + \delta_j) - \boldsymbol{z}_{i,j}^{OT}\right) = 2\left(N\delta_j + \sum_{i=1}^{N} \boldsymbol{x}_{i,j} - \boldsymbol{z}_{i,j}^{OT}\right) , \tag{17}$$

where the first equals is by Equation 16 and noticing that other terms do constants w.r.t. δ_j and the rest is simple calculus. Solving this for δ_j yields the following simple solution:

$$\delta_j = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{z}_{i,j}^{OT} - \boldsymbol{x}_{i,j} = \mu_j^{\boldsymbol{z}^{OT}} - \mu_j^{\text{src}} = \mu_j^{\text{tgt}} - \mu_j^{\text{src}} \tag{18}$$

where the second is just by definition of the mean, and the last is by noticing that the mean of the projected OT samples is equal to the man of the target samples since the projected samples will match the target dataset by construction of the OT

solution. This solution matches the one in the main paper which we recall here:

$$\forall j, [T(\boldsymbol{x})]_j = \begin{cases} x_j + (\mu_j^{\text{tgt}} - \mu_j^{\text{src}}), & \text{if } j \in \mathcal{A} \\ x_j, & \text{if } j \notin \mathcal{A} \end{cases}, \tag{19}$$

Thus showing the optimal delta vector to minimize k-vector transport is exactly the k-sparse mean shift solution.

B. Challenges of Explaining Distribution Shifts and Limitations of Our Method

Distribution shift is a ubiquitous and quite challenging problem. Thus, we believe discussing the challenges of this problem and the limitations of our solution should aid in advancements in this area of explaining distribution shifts.

As mentioned in the main body, as distribution shifts can take many forms, trying to explain a distribution shift is a highly context-dependent problem (i.e., dependent on the data setting, task setting, and operator knowledge). Thus, a primary challenge in developing distribution shift explanations is determining how to evaluate whether a given explanation is valid for a given context. In this work, we hope to introduce the problem of explaining distribution shifts *in general* (i.e. not with a specific task nor setting in mind), therefore we do not have an automated way of measuring whether a given explanation is indeed interpretable. Evaluating explanations is an active area of research (Robnik-vSikonja & Bohanec, 2018; Molnar, 2020; Doshi-Velez & Kim, 2017) with commonalities such as an explanation should be contrastive, succinct, should highlight abnormalities, and should have high fidelity. Instead, we introduce a proxy method that supplies the operator with the PercentExplained and the adjustable k-level of sparse/cluster mappings but leaves the task of validating the explanation up to the operator. We believe developing new shift explanation maps and criteria for specific applications (e.g., explaining the results of experiments run with different initial conditions) is a rich area for future work.

Explaining distribution shifts becomes more difficult when the original data is not interpretable. This typically can take two forms: 1) the raw data *features* are uninterpretable but the samples are interpretable (e.g., a sample from the CelebA dataset (Liu et al., 2015) is interpretable but the pixel-level features are not) or 2) when both the raw data features and samples are uninterpretable (e.g., raw experimental outputs from material science simulations). In the first case, one can use the set of counterfactual pairs method outlined in subsection 3.4 (see Figure 8 for examples with CelebA), however, as mentioned in the main paper, this is less sample efficient than an interpretable transport map. For the second case, if the original features are not interpretable, one must first find an interpretable latent feature space – which is a challenging problem by itself. As seen in Figure 10, it is possible to solve for a semantic latent space and solve interpretable transport maps within the latent space, in this case, the latent space of a VAE model. However, if the meaningful latent features are not extracted, then any transport map within this latent space will be meaningless. In the case of Figure 10, the 3-cluster explanation is likely only interpretable because we know the ground truth and thus know what to look for. As such, this is still an open problem and one we hope future work can improve on.

Additionally, while the PercentExplained metric shows the fidelity of an explanation (i.e. how aligned $T_\sharp(P_{src})$ and P_{tgt} are), we do not have a method of knowing specifically *what* is missing from the explanation. This missing part of the explanation can be considered a "known unknown". For example, if a given T has a PercentExplained of 85%, we know how much is missing, but we do not know what information is contained in the missing 15%. Similarly, when trying to explain an image-based distribution shift with large differences in content (e.g., a dataset with blonde humans and a dataset with bald humans), leveraging existing style transfer architectures (where one wishes to only change the style of an image while retaining as much of the original content as possible) to generate distributional counterfactuals can lead to misleading explanations. This is because explaining image-based distribution shifts might require large changes in content (such as removing head hair from an image), which most style-transfer models are biased against doing. As an example, Figure 8 shows an experiment that translates between five CelebA domains (blond hair, brown hair, wearing hat, bangs, bald). It can be seen that the StarGAN model can successfully translate between stylistic differences such as "blond hair" $\rightarrow$ "brown hair" but is unable to make significant content changes such as "bangs" $\rightarrow$ "bald".

The above issues are mainly problems that affect distribution shift explanations in general, but below are issues specific to our shift explanation method (or any method which similarly uses empirical OT). Since we rely on the empirical OT solution for the sparse and cluster transport (and the percent explained metric), the weaknesses of empirical OT are also inherited. For example, empirical OT, even using the Sinkhorn algorithm with entropic regularization, scales at least quadratically in the number of samples (Cuturi, 2013). Thus, this is only practical for thousands of samples. Furthermore, empirical OT is known to poorly approximate the true population-level OT in high dimensions although entropic regularization can

reduce this problem (Genevay et al., 2019). Finally, empirical OT does not provide maps for new test points. Some of these problems could be alleviated by using recent Wasserstein-2 approximations to optimal maps via gradients of input-convex neural networks based on the dual form of Wasserstein-2 distance (Korotin et al., 2019; Makkuva et al., 2020). Additionally, when using k-cluster maps, the clusters are not guaranteed to be significant (i.e. it might be indiscernible what makes this cluster different than another cluster), and thus if using k-cluster maps on datasets that do not have natural significant clusters (e.g., $P_{src} \sim$Uniform$(0,1)$, $P_{tgt} \sim$Uniform$(1,2)$) an operator might waste time looking for significance where there is none. While this cannot be avoided in general, using a clustering method that is either specifically designed for finding interpretable clusters (Fraiman et al., 2013; Bertsimas et al., 2021) or one which directly optimizes the objective in interpretable transport equation Equation 1 might lead to easier to explanations which are easier to interpret or validate.

C. Experiments on Known Shifts

Here we present additional results on simulated experiments as well as an experiment on UCI "Breast Cancer Wisconsin (Original)" dataset (Mangasarian & Wolberg, 1990). Our goal is to illuminate when to use the different sets of interpretable transport, and how the explanations can be interpreted, where in this case, a ground truth explanation is known. [2]

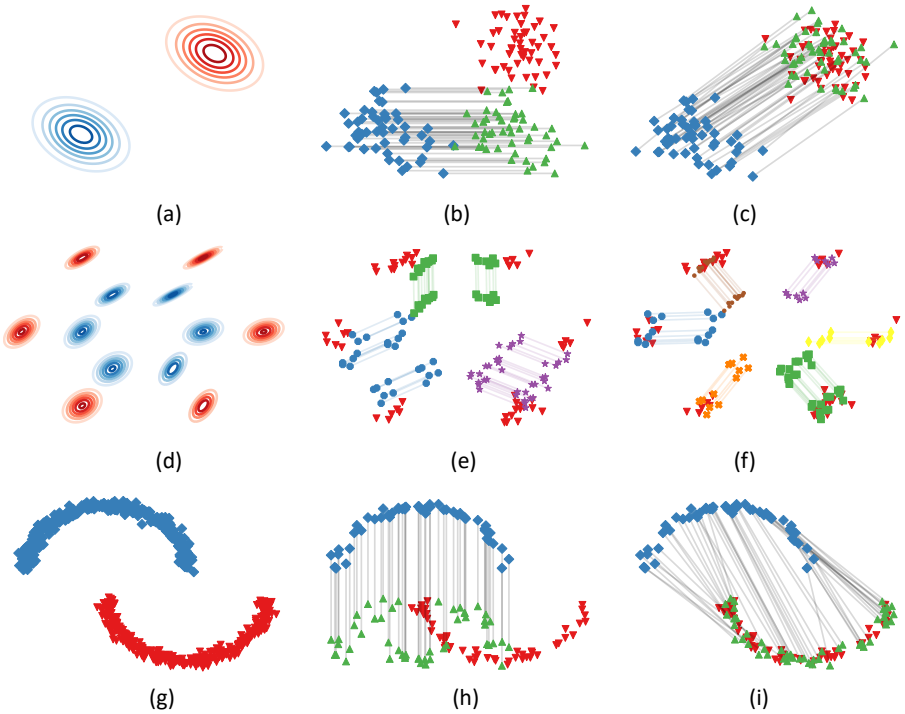


Figure 4: Three toy dataset shift examples showing the advantages of the different shift explanation methods, where a mean shift between Gaussians (top row) can be easily explained using k-sparse vector shifts, a varying mean shift across mixture components of a Gaussian mixture model (middle row) is best explained using k-sparse transport maps, while a complex shift (bottom row) requires a complex feature-wise mapping, such as k-sparse optimal transport, which maximally aligns the distributions as it can perform conditional transport mappings for each sample (as seen by the differing vertical shifts in (h) depending on where the blue sample lies on the horizontal axis), at the expense of interpretability. Each example shows three levels of decreasing interpretability, where the leftmost column shows the original shift (which has maximal interpretability since $k = 0$) from source (blue diamonds) to target (red down arrows), and the rightmost column shows a shift with near-perfect fidelity.

[2]Code to recreate all experiments can be found at https://github.com/inouye-lab/explaining-distribution-shifts.

C.1. Simulated Experiments

In this section we study three toy shift problems: a mean shift between two, otherwise identical, Gaussian distributions, a Gaussian mixture model where each mixture component has a different mean shift, and a flipped and shifted half-moon dataset, as seen in figures (a), (d), and (g) in Figure 4.

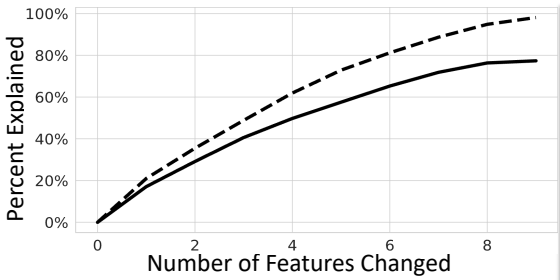
The first case is a mean shift between two, otherwise identical, Gaussian distributions can be easily explained using k-sparse mean shift (as well as vanilla mean shift). We first calculate the OT mapping T_{OT} between the two Gaussian distributions, which has a closed form solution of $T_{OT}(\boldsymbol{x}) = \mu_{tgt} + A(\boldsymbol{x} - \mu_{src})$, where A is a matrix that can be seen as a conversion between the source and target covariance matrices, and because the covariance matrices are identical, A is the identity.

The second toy example of distribution shift is a shifted Gaussian mixture model which represents a case where groups within a distribution shift in different ways. An example of this type of shift could be explaining the change in immune response data across patients given different forms of treatment for a disease. Looking at (d) in Figure 4, it is clear that sparse feature transport will not easily explain this shift. Instead, we turn to cluster-based explanations, where we first find k paired clusters and attempt to show how these shift from P_{src} to P_{tgt}. Following the mean-shift transport of paired clusters approach outlined in subsection 4.3, the $k = 3$ case as seen in the Appendix shows that three clusters can sufficiently approximate the shift by averaging the shift between similar groups. If a more faithful explanation is required, (f) of Figure 4 shows that increasing k to 6 clusters can recover the full shift, i.e. PercentExplained=100, at the expense of being less interpretable (which is especially true in a real-world case where the number of dimensions might be large).

The half-moon example, figure (g) in Figure 4, shows a case where a complex feature-wise dependency change has occurred. This example is likely best explained via feature-wise movement, so will use k-sparse transport. If we follow the approach in subsection 4.2 with our interpretable set as the $\Omega^{(k)}$ and let $k = 1$, we get a mapping that is interpretable, but has poor alignment (see Figure (h) in Figure 4). For this example, we can possibly reject this explanation due to a poor PercentExplained. With the understanding that this shift is not explainable via just one feature, we can instead use a $k = 2$-sparse OT solution. The $k = 2$ case can be seen in (i) of Figure 4 which shows that this approach aligns the distributions perfectly, at the expense of interpretability.

C.2. Explaining Shift in Wisconsin Breast Cancer Dataset

This tabular dataset consists of tumor samples collected by Mangasarian & Wolberg (1990) where each sample is described using nine features which are normalized to integers from $[0, 10]$. We split the dataset along the class dimension and set P_{src} to be the 443 benign tumors and P_{tgt} as the 239 malignant samples. To explain the shift, we used two forms of k-sparse transport, the first being k-sparse mean transport and the second being k-sparse optimal transport. The left of Figure 5 shows that the k-sparse mean shift explanation is sufficient for capturing the 50% of the shift between P_{src} and P_{tgt} using only four features, and nearly 80% of the shift with all 9 features. However, if an analyst requires a more faithful mapping, they can use the k-sparse OT explanation which can recover the full shift, at the expense of the interpretability. The right of Figure 5 shows example explanations that an analyst can use along with their context-specific expertise for determining the main differences between the different tumors they are studying.



2 Features: "Bare Nuclei and Cell Size were shifted by 6.28 and 5.27"

5 Features: "Bare Nuclei, Cell Size, Cell Shape, Normal Nucleoli, and Marginal Adhesion were shifted by 6.28, 5.27, 5.15, 4.6, 4.24"

All Features: "Bare Nuclei, Cell Size, Cell Shape, Normal Nucleoli, Marginal Adhesion, Clump Thickness, Bland Chromatin, Single Epithelial Cell Size, and Mitoses were shifted by 6.28, 5.27, 5.15, 4.6, 4.24, 4.22, 3.89, 3.22, 1.54

Figure 5: A comparison of the performance of k-sparse mean shift explanations (solid line) and k-sparse optimal transport explanations (dashed line) when explaining the shift from the benign tumor samples to malignant tumor samples for the UCI Wisconsin Breast Cancer dataset. On the right are example explanations a human operator would see as they change the level of interpretability during k-sparse mean shift explanations (where "All Features" is the baseline full mean shift explanation).

C.3. Counterfactual Example Experiment to Explain a Multi-MNIST shift

As mentioned in subsection 3.4, image-based shifts can be explained by supplying an operator with a set of distributional counterfactual images with the notion that the operator would resolve which semantic features are distribution-specific. Here we provide a toy experiment (as opposed to the real-world experiment seen in subsection 3.4) to illustrate the power of distributional counterfactual examples. To do this, we apply the distributional counterfactual example approach to a Multi-MNIST dataset where each sample consists of a row of three randomly selected MNIST digits (Deng, 2012) and is split such that P_{src} consists of all samples where the middle digit is even and zero and P_{tgt} is all samples where the middle digit is odd, as seen in Figure 6.

Figure 6: A grid of 25 raw samples from each domain (left is P_{src} and right is P_{tgt}). Even for the relatively simple shift seen in the Shifted Multi-MNIST dataset, it may be hard to tell what is different between the two distributions by just looking at samples (without knowing the oracle shift). Each sample in this dataset contains three MNIST digits along a diagonal and the domain label corresponds to the evenness of the middle MNIST digit (where P_{src} contains even middle digits and P_{tgt} contains odd middle digits).

Algorithm 3 Generating distributional counterfactuals using DIVA

Input: $\boldsymbol{x}_1 \sim D_1$, $\boldsymbol{x}_2 \sim D_2$, model
$z_{y_1}, z_{d_1}, z_{residual_1} \leftarrow$ model.encode($\boldsymbol{x}_1$)
$z_{y_2}, z_{d_2}, z_{residual_2} \leftarrow$ model.encode($\boldsymbol{x}_2$)
$\hat{\boldsymbol{x}}_{1\to2} \leftarrow$ model.decode($z_{y_1}, z_{d_2}, z_{residual_1}$)
$\hat{\boldsymbol{x}}_{2\to1} \leftarrow$ model.decode($z_{y_2}, z_{d_1}, z_{residual_2}$)
Output: $\hat{\boldsymbol{x}}_{1\to2}$, $\hat{\boldsymbol{x}}_{2\to1}$

To generate the counterfactual examples, we use a Domain Invariant Variational Autoencoder (DIVA) (Ilse et al., 2020), which is designed to have three independent latent spaces: one for class information, one for domain-specific information (or in this case, distribution-specific information), and one for any residual information. We trained DIVA on the Shifted Multi-MNIST dataset for 600 epochs with a KL-β value of 10 and latent dimension of 64 for each of the three sub-spaces. Then, for each image counterfactual, we sampled one image from the source and one image from the target and encoded each image into three latent vectors: z_y, z_d, and $z_{residual}$. The latent encoding z_d was then "swapped" between the two encoded images, and the resulting latent vector set was decoded to produce the counterfactual for each image. This process is detailed in Algorithm 2 below. The resulting counterfactuals can be seen in Figure 7 where the middle digit maps from the source (i.e., odd digits) to the target (i.e., even digits) and vice versa while keeping the other content unchanged (i.e., the top and bottom digits).

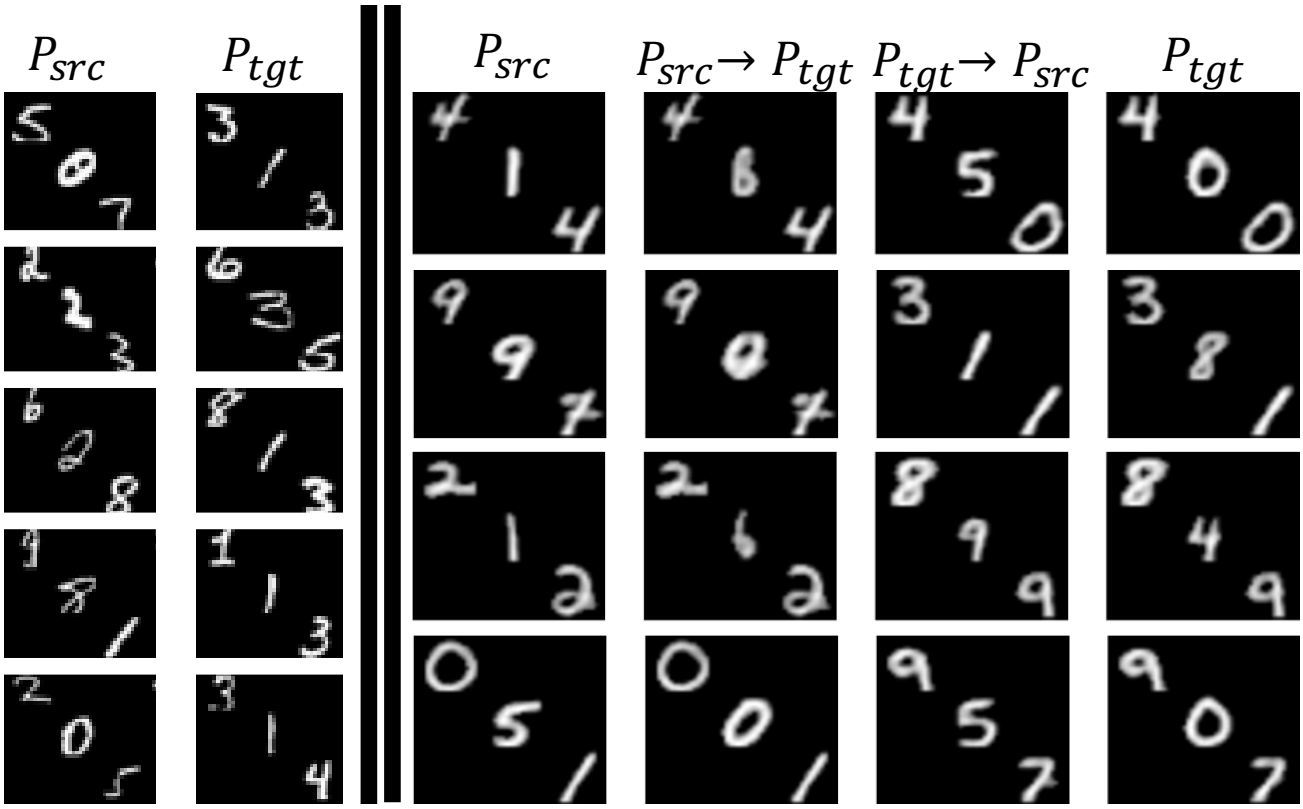


Figure 7: A comparison of the baseline grid of unpaired source and target samples (left) and counterfactual pairs (right) which show how counterfactual examples can highlight the difference between the two distributions. For each image, the top left digit represents the class label, the middle digit represents the distribution label (where P_{src} only contains even digits and zero and P_{tgt} has odd digits), and the bottom right digit is noise information and is randomly chosen. The second, third columns show the counterfactuals from $P_{src} \rightarrow P_{tgt}$ and $P_{tgt} \rightarrow P_{src}$, respectively. Hence we can see under the push forward of each image the "evenness" of the domain digit changes while the class and noise digits remain unchanged.

C.4. Using StarGAN to Explain Distribution Shifts in CelebA

Here we apply the distributional counterfactual approach seen in subsection 3.4 to the CelebA dataset (Liu et al., 2015), which contains over 200K images of celebrities, each with 40 attribute annotations. We split the original dataset into 5 related sets, P_1="blonde hair", P_2="brunette hair", P_3="wearing hat", P_4="bangs", P_5="bald". These five sets were chosen as they are related concepts (all related to hair) yet mostly visually distinct. Although there are images with overlapping attributes, such as a blonde/brunette person with bangs, these are rare and naturally occurring, thus they were not excluded.

We trained a StarGAN model (Choi et al., 2018) to generate distributional counterfactuals following the same approach seen in subsection 3.4. The result of this process can be seen in Figure 8, where we can see the model successfully translating "stylistic" parts of the image such as hair color. However, the model is unable to translate between distributions with larger differences in "content" such as removing hair when translating to "bald". This highlights a difference between I2I tasks such as style transfer (where one wishes to bias a model to only change the style of an image while retaining as much of the original content as possible) the mappings required for explaining image-based distribution shifts, which might require large changes in content (such as adding a hat to an image).

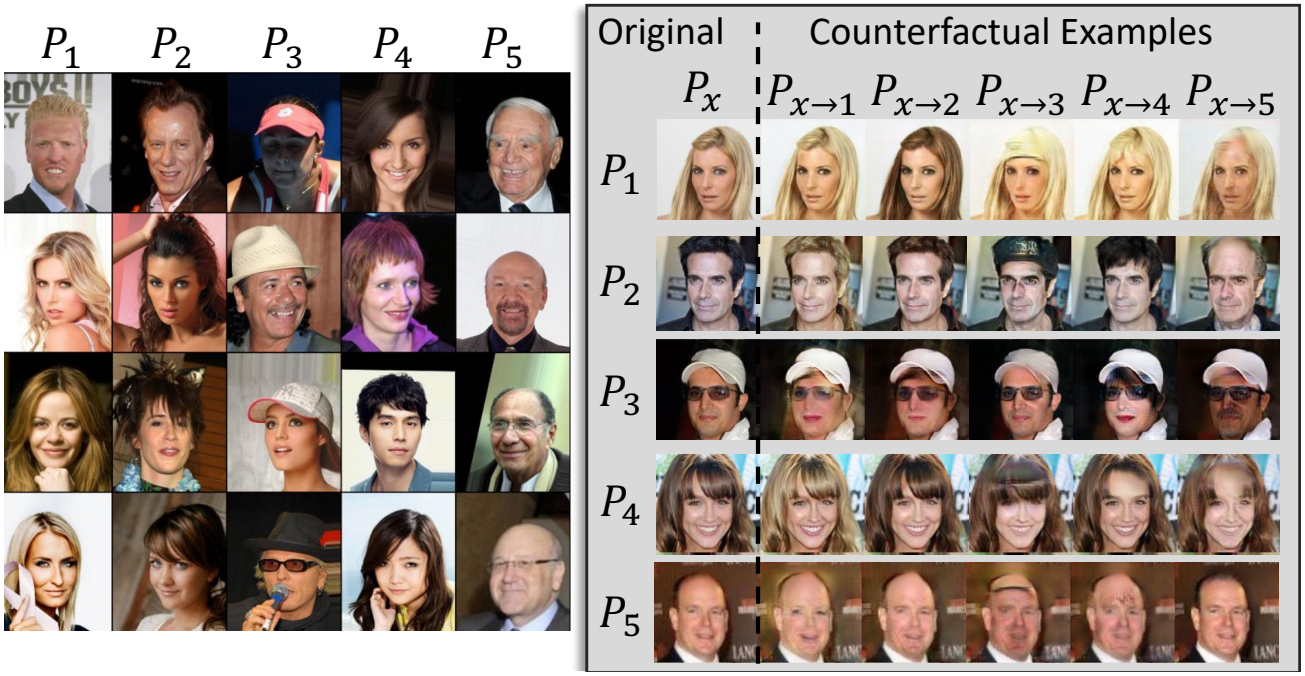


Figure 8: StarGAN is able to adequately translate between distributions with similar content but different style (e.g., $P_1 \to P_2$), however, when transporting between distributions with different content (e.g., "no hat" $\to P_3$) the I2I model is unable to properly capture the shift. This is likely due to the model being biased to only change the *style* of the image, while maintaining as much *content* as possible. The figure breakdown is similar to Figure 3 with the baseline method of unpaired samples on the left and paired counterfactual images on the right, where here P_1="blonde hair", P_2="brunette hair", P_3="wearing hat", P_4="bangs", P_5="bald".

D. Explaining Shifts in Images via High-Dimensional Interpretable Transportation Maps

If $\boldsymbol{x}$ is an image with domain $\mathbb{R}^{D>>1}$, then any non-trivial transportation map in this space is likely to be hard to optimize for as well as uninterpretable. However, if P_{src}, P_{tgt} can be expressed on some *interpretable* lower dimensional manifold which is learned by some manifold-invertible function $g : \mathbb{R}^D \to \mathbb{R}^{D'}$ where $D < D'$, we can project P_{src}, P_{tgt} onto this latent space and solve for an interpretable mapping such that it aligns the distributions in the latent space, $P_{T(g(\boldsymbol{x}))} \approx P_{g(\boldsymbol{y})}$. Note, in practice, an encoder-decoder with an interpretable latent space can be used for g, however, requiring g to be exactly invertible allows for mathematical simplifications, which we will see later. For explainability purposes, we can use g^{-1} to re-project $T(g(\boldsymbol{x}))$ back to $\mathbb{R}^D$ in order to display the transported image to an operator. With this, we can define our set of high dimensional interpretable transport maps: $\Omega_{\text{high-dim}} :- \left\{T : T = g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right), \tilde{T} \in \Omega^{(k)}, g \in \mathcal{I}\right\}$ where $\Omega^{(k)}$ is the set of k-interpretable mappings (e.g., k-sparse or k-cluster maps) and $\mathcal{I}$ is the set of invertible functions with an interpretable (i.e. semantically meaningful) latent space.

Looking at our interpretable transport problem:

$$\underset{T \in \Omega_{\text{high-dim}}}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c(\boldsymbol{x}, T(\boldsymbol{x}))\right] + \lambda\phi(P_{T(\boldsymbol{x})}), P_{\boldsymbol{y}}) \tag{20}$$

Although our transport is now happening in a semantically meaningful space, our transportation cost is still happening in the original raw pixel space. This is undesirable since we want a transport cost that penalizes large semantic movements, even if the true change in the pixel space is small (e.g., a change from "dachshund" to "hot dog" would be a large semantic movement). We can take a similar approach as before and instead calculate our transportation cost in the g space. This logic can similarly be applied to our divergence function ϕ (especially if ϕ is the Wasserstein distance, since this term can be seen as the residual shift not explained by T). Thus, calculating our cost and alignment functions within the latent space gives us:

$$\underset{g \in \mathcal{I}, \tilde{T} \in \Omega^{(k)}}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c\left(g(\boldsymbol{x}), \tilde{T}\left(g(\boldsymbol{x})\right)\right)\right] + \lambda\phi(P_{\tilde{T}(g(\boldsymbol{x}))}, P_{g(\boldsymbol{y})}) \tag{21}$$

This formulation has a critical problem however. Since we are jointly learning our representation g and our transport map T, a trivial solution for the above minimization is for g to map each point to an arbitrarily small space such that the distance between any two points $c(g(\boldsymbol{x}), g(\boldsymbol{y})) \approx 0$, thus giving us a near zero cost regardless of how "far" we move points. To avoid this, we can use pre-defined image representation function h, e.g., the latter layers in Inception V3, and calculate pseudo-distances between transported images in this space. Because h expects an image as an input, we can utilize the invertibility of g and perform our cost calculation as: $c\left(h(\boldsymbol{x}), h\left(g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right)\right)\right)$, or more simply, $c_h\left(\boldsymbol{x}, T(\boldsymbol{x})\right)$, where $T = g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right)$. Similar to the previous equation, we also apply this h pseudo-distance to our divergence function to get ϕ_h. With this approach, we can still use g to jointly learn a semantic representation which is specific to our source and target domains (unlike h which is trained on images in general, e.g., ImageNet) and an interpretable transport map $\tilde{T}$ within g's latent space. This gives us:

$$\underset{g \in \mathcal{I}, T \in \Omega}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c_h\left(\boldsymbol{x}, T(\boldsymbol{x})\right)\right] + \lambda \phi_h(P_{T(\boldsymbol{x})}, P_{\boldsymbol{y}}) \tag{22}$$

While the above equation is an ideal approach, it can be difficult to use standard gradient approaches to optimize over in practice due it being a joint optimization problem and any gradient information having to first pass through h which could be a large neural network. To simplify this, we can optimize $\tilde{T}$ and g separately. With this, we can first find a g which properly encodes our source and target distributions into a semantically meaningful latent space, and then find the best interpretable transport to align the distributions in the fixed latent space. The problem can be even further simplified by setting the pre-trained image representation function h to be equal to the pretrained g, since the disjoint learning of g and $\tilde{T}$ removes the shrinking cost problem. By setting $h := g$, we can see that $c\left(h(\boldsymbol{x}), h \circ g^{-1} \circ \tilde{T} \circ g(\boldsymbol{x})\right) = c\left(g(\boldsymbol{x}), \tilde{T} \circ g(\boldsymbol{x})\right) = c_g(\boldsymbol{x}, \tilde{T}(\boldsymbol{x}))$, which simplifies Equation 22 back to our interpretable transport problem, Equation 20, where g is treated as a pre-processing step on the input images:

$$\underset{T \in \Omega}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c(g(\boldsymbol{x}), g\left(T(\boldsymbol{x})\right))\right] + \lambda \phi_g(P_{T(\boldsymbol{x}))}, P_{\boldsymbol{y}}) \tag{23}$$

Another way to simplify Equation 22 is to relax the constraint that g is manifold-invertible and instead use a pseudo-invertible function such as an encoder g and decoder g^+ structure where g^+ is a pseudo-inverse to g such that $g^+(g(\boldsymbol{x})) \approx \boldsymbol{x}$. This gives us:

$$\begin{aligned}\underset{\tilde{T} \in \tilde{\Omega}, g, g^+}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c_h\left(\boldsymbol{x}, g^+(\tilde{T}(g(\boldsymbol{x})))\right)\right] &+ \lambda_{Fid}\ \phi_h(P_{g^+(\tilde{T}(g(\boldsymbol{x})))}, P_{\boldsymbol{y}}) \\ &+ \lambda_{Recon}\ \mathbb{E}_{\frac{1}{2}P_{src}+\frac{1}{2}P_{tgt}}\left[L\left(\boldsymbol{x}, g^+(\tilde{T}(g(\boldsymbol{x})))\right)\right]\end{aligned} \tag{24}$$

where $L(\boldsymbol{x}, \cdot)$ is some reconstructive-loss function.

D.1. Explaining a Colorized-MNIST shift via High-dimensional Interpretable Transport

In this section we present a preliminary experiment showing the validity of our framework for explaining high-dimensional shifts. The experiment consists of using k-cluster maps to explain a shift in a colorized-version of MNIST, where the source environment is yellow/light red digits with a light grayscale background color (i.e. light gray) and the target environment consists of darker red digits and/or a darker grayscale background colors. Like the lower dimensional experiments before, our goal is to test our method on a shift where the ground truth is known and thus the explanation can validated against. We follow the framework presented in Equation 23, where the fixed g is a semi-supervised VAE (Siddharth et al., 2017) which is trained on a concatenation of P_{src} and P_{tgt}. Our results show that k-cluster transport can capture the shift and explain the shift, however, we suspect the given explanation is interpretable because the ground truth is already known. Our hope is that future work will improve upon this framework by better finding a latent space which is interpretable and disentangled, leading to better latent mappings, and thus better high-dimensional shift explanations.

Data Generation The base data is the 60,000 grayscale handwritten digits from the MNIST dataset (Deng, 2012). We first colored each digit by copying itself along the red and green channel axes with an empty blue channel, yielding an initial dataset of yellow digits. We then randomly sampled 60,000 points from a two-dimensional Beta distribution with shape

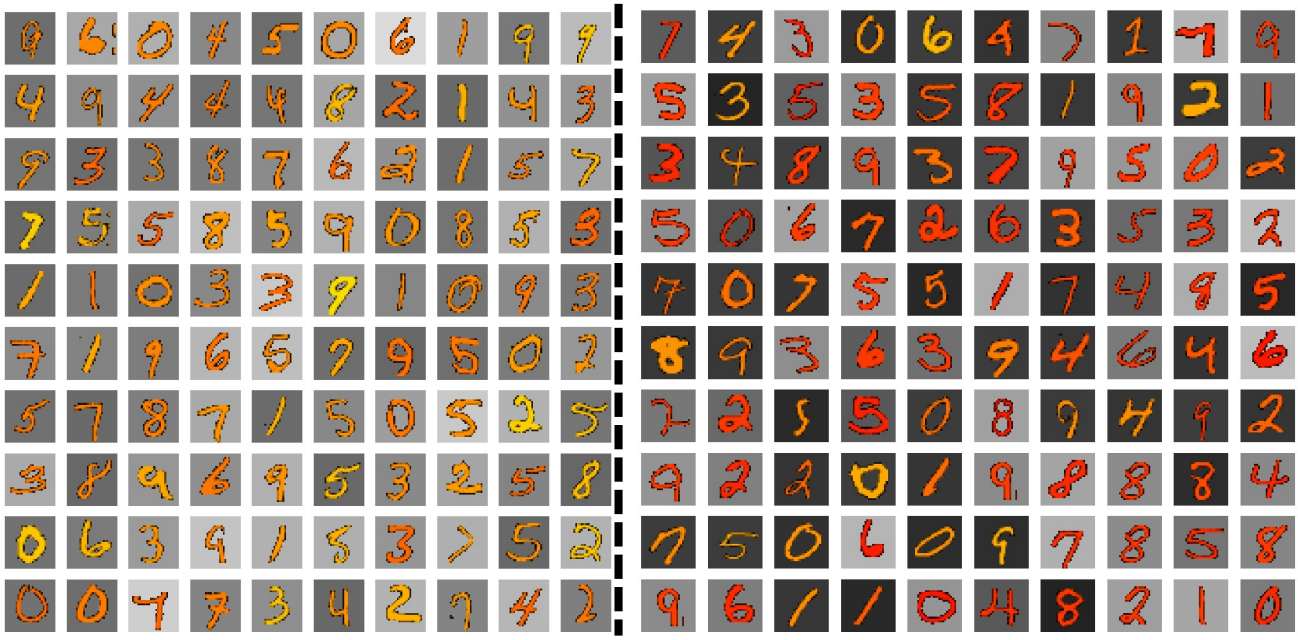


Figure 9: The left figure shows samples from the source environment which has lighter digits and backgrounds while the right figure shows the target environment which has darker digits and/or darker backgrounds

parameters, $\alpha = \beta = 5$. The first dimension of our Beta distribution represented how much of the green channel would be visible per sample meaning low values would result in a red image, while high values would result in a yellow image. The second dimension of our Beta distribution represented how white vs. black the background of the image would be, where $0 :=$ black background and $1 :=$ white background.

Specifically, the data was generated as follows. With $\boldsymbol{x}_{raw}$ representing a grayscale digit from the unprocessed MNIST dataset, a mask of representing the background was calculated $\mathbf{m} = \boldsymbol{x}_{raw} \leq 0.1$, where any pixel value below 0.1 is deemed to be the background (where each pixel value $\in [0, 1]$). Then, the foreground (i.e. digit) color was created $\boldsymbol{x}_{digit-color} = [(1 - \mathbf{m}) \cdot \boldsymbol{x}_{raw},\ b_1 \cdot (1 - \mathbf{m}) \cdot \boldsymbol{x}_{raw}, \mathbf{0}]$, where $\mathbf{0}$ is a zero-valued matrix matching the size of $\boldsymbol{x}_{raw}$ and $b_1 \sim \text{Beta}(\alpha, \beta)$. The background color was calculated via $\boldsymbol{x}_{back-color} = [b_2 \cdot \mathbf{m} \cdot \boldsymbol{x}_{raw},\ b_2 \cdot \mathbf{m} \cdot \boldsymbol{x}_{raw},\ b_2 \cdot \mathbf{m} \cdot \boldsymbol{x}_{raw}]$. Then $\boldsymbol{x}_{colored} = \boldsymbol{x}_{digit-color} + \boldsymbol{x}_{back-color}$, which results in a colorized MNIST digit with a stochastic foreground and background coloring.

The environments were created by setting the source environment to be any images where $b_1 \geq 0.4$ and $b_2 \geq 0.4$, i.e. any colorized digits that had over 40% of the green channel visible *and* a background at least 40% white, and the target environment is all other images. Informally, this split can be thought of as three sub-shifts: a shift which is only reddens the digit, a second shift which only a darkens the background, and a third shift which is both a digit reddening and background darkening. The environments can be seen in Figure 9.

Model To encode and decode the colored images, we used a semi-supervised VAE (SSVAE) (Siddharth et al., 2017). The SSVAE encoder consisted of an initial linear layer with input size of $3 \cdot 28 \cdot 28$ and a latent size of 1024. This was then multi-headed into a classification linear layer of size 1024 to 10, and for each sample with a label, digit label classification was performed on the output of this layer. The second head of the input layer was sent to a style linear layer of size 1024 to 50 which is to represent the style of the digit (and is not used in classification). The decoder followed a typical VAE decoder approach on a concatenation of the classification and style latent dimensions. The SSVAE was trained for 200 epochs on a concatenation of both P_{src} and P_{tgt} with 80% of the labels available per environment, and a batch size of 128 (for training details please see (Siddharth et al., 2017)). The transport mapping was then found on the saved lower-dimensional embeddings.

Shift Explanation Results Given the shift is divided into three main sub-shifts, we used $k = 3$ cluster maps to explain the shift. We followed the approach given in Equation 23, where the three cluster labels and transport were found in the 60

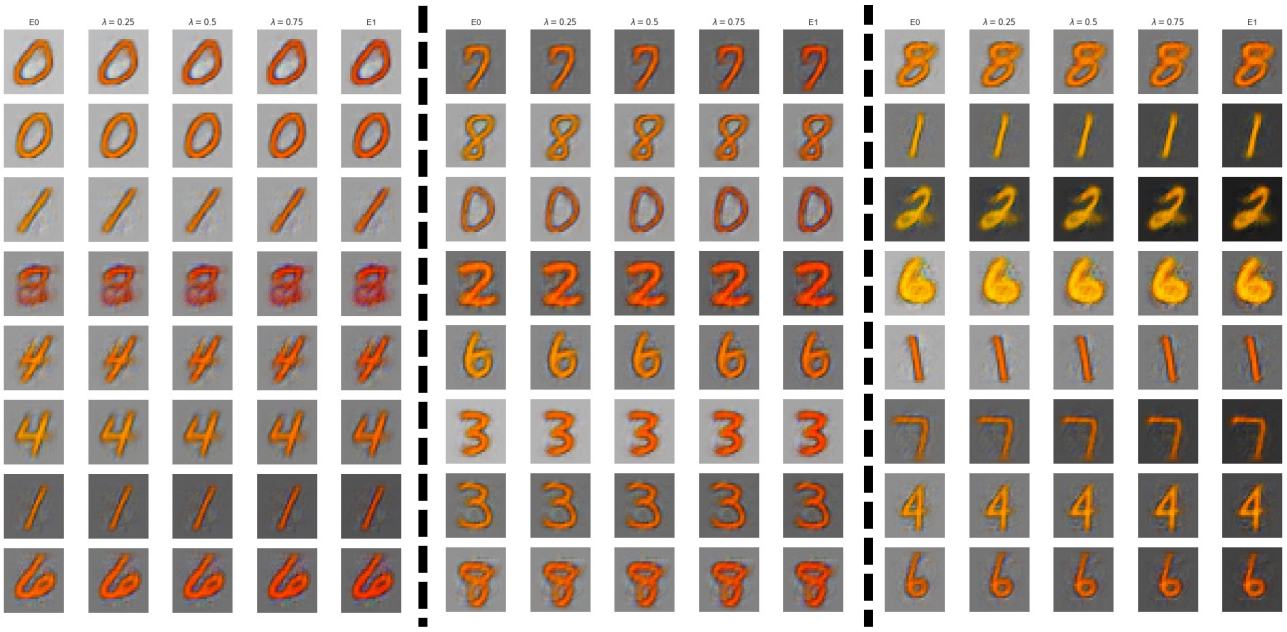


Figure 10: The linear interpolation explanations for the three clusters where the left cluster *seems* to explain the darkening digit shift, the right-most figure explains the shift which darkens the background, and the middle cluster explains the case where both digit and background darkens. For each cluster, the left-most digit $\boldsymbol{x}$ is the reconstruction of original encoding from the source distribution, the right-most digit is the cluster-based push-forward of that digit $T(\boldsymbol{x})$, and the three middle images are reconstructions of a linear interpolations $\lambda \cdot \boldsymbol{x} + (1 - \lambda) \cdot T(\boldsymbol{x})$ with $\lambda \in \{0.25, 0.5, 0.75\}$.

dimensional latent space using the algorithm given in Algorithm 1. Since our current approach is not able to find a latent space with disentangled and semantically meaningful axes, we cannot use the mean shift information per cluster as the explanation itself (as it is meaningless if the space is uninterpretable). Instead, we provide an operator with m samples from our source environment and the linear interpolation to the samples' push-forward versions under the target environment, for each cluster. The goal is for the operator to discern the meaning of each cluster's mean shift by finding the invariances across the m linear interpolations. The explanations can be seen in Figure 10.

The linear interpolations from the first cluster (the left of Figure 10) seem to show a darkening of the source digit, while keeping the background relatively constant. The third cluster (right-most side of the figure) represents the situation where only the background is darkened but the digit is not. Finally, the third cluster seems to explain the sub-shift where both the background and the digit are darkened. However, the changes made in the figures are quite faint, and without *a priori* knowledge of the shift it is possible that this could be an insufficient explanation. As mentioned in Appendix D, this could be improved by finding a disentangled latent space with semantically meaningful dimensions, better approximating high dimensional empirical optimal transport maps, jointly finding a representation space and transport map like in Equation 23, and more; however, these advancements are out of scope for this work. We hope that this current preliminary experiment showcases the validity of using transportation maps to explain distribution shifts in images and inspires future work to build upon this foundation.