Towards Explaining Image-Based Distribution Shifts

Sean Kulinski, David I. Inouye
School of Electrical and Computer Engineering
Purdue University
{skulinsk, dinouye}@purdue.edu

Abstract

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 such distribution shifts is critical for examining and hopefully mitigating the effect of such a shift. Most prior work has focused on either natively handling distribution shift (e.g., Domain Generalization) or merely detecting a shift while assuming any detected shift can be understood and handled appropriately by a human operator. For the latter, we hope to aid in these manual mitigation tasks by explaining the distribution shift to an operator. To this end, we suggest two methods: providing a set of interpretable mappings from the original distribution to the shifted one or providing a set of distributional counterfactual examples. We provide preliminary experiments on these two methods, and discuss important concepts and challenges for moving towards a better understanding of image-based distribution shifts.

1. Introduction

Most real-world environments are constantly changing and understanding how a specific operating environment has changed is crucial to making decisions respective to such a change. Such a change might be a new data distribution seen in deployment which causes a machine learning model to begin to fail. 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 by providing an explanation of such a change.

This ubiquitous phenomena 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 on *understanding* or *characterizing* a detected distribution shift. A naïve baseline in analyzing an image-based distribution shift is to compare a grid of samples from the original, i.e., *source*, distribution to a grid of samples from the new, i.e., *target*, distribution. However, due to the complexity of image-based shifts, this approach can be uninterpretable or even misleading to an operator (e.g., the left parts of Figure 1 and Figure 2).

Therefore, we propose two preliminary methods for explaining image-based distribution shifts and discuss open challenges. The first is a novel framework which provides an operator with interpretable mappings which shows how latent features have changed or how latent groups have shifted between the distributions. The second approach is similar to that of unpaired Image-to-Image Translation (I2I) [14] such as CycleGAN [24], and explains the shift to the operator as pairs of a real example and its corresponding counterfactual example. These counterfactuals are generated by mapping samples from one domain to the other domain such that the distributions become indistinguishable. We summarize our contributions as follows:

- We introduce high-dimensional interpretable transport maps for explaining image-based shifts if an interpretable latent space is known.
- We also leverage prior I2I work to explain image-based distribution shifts via counterfactual examples if an interpretable latent space is unavailable.
- We provide preliminary results and interpretations.
- We discuss open questions for explaining image-based distribution shifts.

2. Explaining Image Distribution Shifts via Transportation 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 $T(\boldsymbol{x}) \sim P_{tgt}$. Additionally, if an interpretable representation of the map

T can be formed, this representation can be provided to an operator to aid in understanding and reacting to shifts more effectively. However, an interpretable representation likely requires interpretable (latent) features, which may not be available for some image domains. In this case, we can represent the map by merely showing pairs of inputs $\boldsymbol{x}$ and "counterfactual" outputs $T(\boldsymbol{x})$. Therefore, we define a shift explanation to be: *a (possibly interpretable) transport map T that maps a source distribution P_{src} onto a target distribution P_{tgt} such that $T_\sharp P_{src} \approx P_{tgt}$.*

2.1. Interpretable Transportation Maps

In order to find such a mapping between distributions, it is natural to look to Optimal Transport (OT) due to it allowing for a rich geometric structure on the space of distributions and having extensive prior work in this field [1, 5, 15, 21]. An OT mapping is originally defined by Monge [15, 22] as a method of aligning two distributions in a minimal cost way given a transport cost function c. To find interpretable transport maps, we build upon the OT framework by restricting the candidate transport maps to belong to a set of user-defined interpretable mappings Ω. Additionally we use a Lagrangian relaxation on the full alignment constraint seen in OT, giving us an *Interpretable Transport* mapping T_{IT}:

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

where $\phi(\cdot,\cdot)$ is a divergence function, which, unless otherwise stated, is assumed to be the squared Wasserstein-2 metric, W_2^2.

An example of an set of interpretable mappings Ω is k-cluster mappings. Where 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 j^{th} column of Δ. For another set of interpretable mappings (k-sparse transport) and methods for solving for these mappings in practice, please see section Appendix C.

In order to find interpretable transport mappings for high dimensional spaces like images, we can project P_{src} and P_{tgt} onto an *interpretable* latent space (e.g., a space which has disentangled and semantically meaningful dimensions) which is learned by some (pseudo-)invertible function $g : \mathbb{R}^d \to \mathbb{R}^k$ where $k < d$ (e.g., an autoencoder). Then we can solve for an interpretable mapping such that it aligns the distributions in the latent space, $P_{T(g(\boldsymbol{x}))} \approx P_{g(\boldsymbol{y})}$. For counterfactual purposes, we can use g^{-1} to project $T\left(g(\boldsymbol{x})\right)$ 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: $\Omega_{\text{high-dim}} \coloneqq \left\{T : T = g^{-1}\left(\tilde{T}\left(g(\boldsymbol{x})\right)\right), \tilde{T} \in \Omega, g \in \mathcal{I}\right\}$ where Ω a the set of interpretable mappings and $\mathcal{I}$ is the set of (pseudo-)invertible functions with a interpretable (i.e., semantically meaningful) latent space. Given an interpretable $g \in \mathcal{I}$, we define our problem as:

$$\underset{\tilde{T} \in \Omega^{(k)}}{\arg\min}\ \mathbb{E}_{P_{src}}\left[c\left(g(\boldsymbol{x}), \tilde{T}\left(g(\boldsymbol{x})\right)\right)\right] + \lambda\phi(P_{\tilde{T}(g(\boldsymbol{x}))}, P_{g(\boldsymbol{y})}) \tag{2}$$

which results in an interpretable map $\tilde{T}$ which approximately shows how images from P_{src} shifted to P_{tgt} in a semantically meaningful way (e.g., how the H&E staining in histopathology images changes across hospitals).

2.2. Counterfactuals via Unpaired Image-to-image Translation

In some cases, a shift cannot be expressed by an interpretable mapping function because an interpretable latent space is not known. Thus, we can remove the interpretability constraint, 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 [14]. Once a 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 the interpretable mapping approach.

3. Experiments

In this section we provide preliminary results showing the advantages and shortcomings of explaining shifts via interpretable transportation maps and via counterfactual pairs. We begin with explaining a shifted Color MNIST dataset via cluster-based transportation maps using a semi-supervised VAE [18]. Next, we use StarGAN [4] to generate counterfactual examples to explain the shift in histopathology images across five hospitals as seen in the Stanford Wilds [9] variant of the Camelyon17 dataset [2]. [1]

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

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

This experiment consists of using k-cluster maps to explain a shift in a colorized-version of MNIST, where the source environment has more yellow digits with a light gray background while the target environment consists of more red digits and/or darker gray backgrounds. The data is created by randomly red/yellow colorizing the foreground and grayscale coloring the background of 60,000 grayscale MNIST digits [6]. The source distribution P_{src} is set to be be any images where colorized digits that had over 40% of the green channel visible (thus yielding a yellow color) *and* a background at least 40% white, and the target environment P_{tgt} is all other images. Informally, this split can be thought of as three heterogeneous sub-shifts: a shift which is only reddens the foreground digit, a second shift which only a darkens the background, and a third shift which both reddens the digit reddening and darkens the background. The environments can be seen in Figure 3 in the Appendix.

We follow the framework presented in Equation 2, where g is a semi-supervised VAE [18] with a latent dimension of 50. 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 and otherwise followed the training details in [18]. To explain the shift, we use Algorithm 1 in the appendix to learn $k = 3$ cluster maps because there are 3 subshifts.

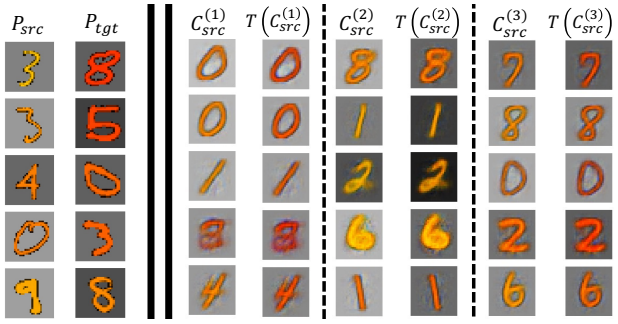

Figure 1. The baseline of unpaired source and target samples (left) is unable to distinguish between the three subshifts. Our cluster-based transport (right) separates the shift into 3 subshifts: $C^{(1)}$ clearly reddens the digit color but maintains the background color, $C^{(2)}$ clearly darkens the background color but maintains the digit color, and $C^{(3)}$ changes both the digit color and background color.

While the cluster map is inherently simple because each map merely translates points by a constant vector, the latent features are not disentangled into semantically meaningful features. Thus, to represent the cluster map, we merely show input and output pairs for each cluster map. The goal is for the operator to discern the meaning of each cluster's shift by finding the invariances for each cluster. The cluster based explanations can be seen in Figure 1. Our preliminary results demonstrate that k-cluster transport can explain this heterogeneous shift by separating at least two distinct shifts in the data. However, we acknowledge that this is a relatively simple example and expect more work will be needed to improve this idea for real-world image shifts.

3.2. Explaining Shifts in H&E Images Across Hospitals via Counterfactual Examples

This experiment explores the alternative for explaining image-based distribution shifts by supplying an operator with a set of translated images (i.e., a set of images from the source distribution which have been altered to look like they belong to the target distribution), with the notion that the operator would resolve which semantic features are distribution-specific. We apply this approach the Camelyon17 dataset [2] which is a real-world distribution shift dataset that consists of whole-slide histopathology images from five different hospitals. We use the Stanford WILDS [9] 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 can be seen in Figure 2.

To generate the counterfactual examples, we treat each hospital as a domain and train a StarGAN model [4] to translate between each domain. For training, we followed the original training approach seen in [4], with the exception that we perform no center cropping. After training, we can generate image counterfactual examples via 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 in the right-hand side of Figure 2. It can be seen that the StarGAN model captures the different staining characteristics across the hospitals. For example, hospital 1 (P_1) consists of mostly light staining and thus transporting to this domain usually involves a lightening of image while P_3 seems to have more hematoxylin stain thus leading to deeper purple images when pushing onto this domain. We can also see that the model tends to respect the content of the image where patches which contain tumor cells (e.g., the P_5 sample on the right-hand side) still contain tumor cells in the counterfactual cases and likewise for lymphocyte cells (e.g., the P_4 sample on right-hand side).

4. Open Questions for Explaining Image-based Distribution Shifts

In this section, we introduce a series of open questions which we hope will help move towards developing the foundations for explaining image-based distribution shifts including defining exemplar tasks, metrics, datasets, and

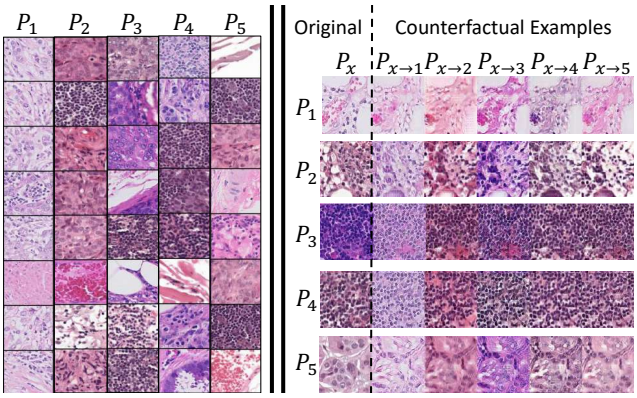

Figure 2. The baseline method of unpaired samples (left) which requires many samples to begin to understand the differences across the hospitals domains (represented as P_1, P_2, ...). Our explanation approach (right) of showing paired counterfactual images translated between the hospital domains (where the (i, j) row, column pair represents the pushforward of the i$^{\text{th}}$ domain onto the j$^{\text{th}}$ domain) quickly makes it clear how the staining/coloring differs across the hospital domains.

baseline methods. We begin with introducing tasks where an operator would need to *understand* a distribution shift and give criteria for finding exemplar datasets which can serve as benchmarks for the tasks. Then we discuss possible other approaches for explaining distribution shift and close with suggesting criteria for evaluating and comparing such methods.

We (non-exhaustively) envision several possible tasks: **Knowledge discovery** - This would entail helping an operator extract knowledge by characterizing the differences between distributions (e.g., finding important differences in nanostructure imaging with different experimental conditions), and would focus on complex distribution shifts that would not be easy to understand using conventional visualization or dataset inspection tools. **Post-hoc explanations of model failure due to shifts** - This would involve finding the qualitative differences between the training environment and this new testing environment that caused the model to fail. It would help an operator answer the question: Can we determine how to alleviate this problem? Should we collect more labeled data, adjust the instrument, or robustify the model? **Detecting adversarial shifts** This would help an operator determine if the distributional changes are due to benign effects or due to an adversary (i.e., an enemy compromises a surveillance camera). Due to the highly context-dependent nature of distribution shift, it would be beneficial to have exemplar datasets on which to train and evaluate methods for each of these tasks. Ideally, these distribution shift examples would be complex distribution shifts—not something that can be easily explained by a simple plot or by looking at the difference in mean statistics—, have real-world use cases where understanding the distribution shift is important, and has some form of known oracle explanation(s) that could be used to validate a predicted explanation against.

In this paper we introduced a novel way for explaining image-based distribution shifts via interpretable transport maps; however, there are other ways characterize and explain an image-based distribution shift. For example, we also discuss and show how image translation works can be used to explain distribution shift via providing an operator with sets of counterfactual pairs. However, we are not sure if the current work in I2I can directly be applied to explain distribution shifts. For example, the problem of style-transfer focuses on transferring the "style" of an image to between two domains while keeping the "content" constant, but what is considered "content" likely needs to be specified by an operator for their specific context in order to be directly actionable (e.g., ensuring road features are considered constant when analyzing human-driving data). Another approach would be to find a causal model of the semantic content between the two distributions, and characterizing the causal differences between them (e.g., the approach of [3] applied to images). In addition to finding methods for explaining image-based distribution shifts, we need ways to evaluate and compare methods. For transport maps, we suggest that a natural metric is to determine how well the transported source distribution aligns with the target distribution via distributional divergences such as Wasserstein distance or KL divergence. However, the interpretability or actionability of a shift explanation is more challenging to define. A proxy method for evaluating this would likely be task specific (but ideally not dataset specific) and should not require expensive human evaluation. For mapping-based methods, the measurement of interpretability could be a function of the complexity of the mapping, however, how to systematically measure the interpretability of counterfactual approaches is currently unclear.

5. Conclusion

In this paper, we introduced a novel framework for explaining image-based distribution shift using transport maps T between a source and target distribution. If a semantically meaningful latent space is known, we can constrain T to be relatively simple. If a meaningful latent space is unavailable, we show how prior image-to-image translation work can explain such shifts via sets of counterfactual examples. We demonstrate both approaches on two distribution shift examples. We then initialized a discussion which hopefully will lead to a better foundation of explaining image distribution shifts. We ultimately hope this work lays the groundwork explaining and thus understanding image distribution shifts.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, 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. 2, 3

[3] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1666–1674. PMLR, 13–15 Apr 2021. 4, 6

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 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. 2, 3

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 2

[6] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 3, 10

[7] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 6

[8] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 11

[9] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 2, 3, 6

[10] Sean Kulinski, Saurabh Bagchi, and David I Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in Neural Information Processing Systems*, 33, 2020. 6

[11] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 6

[12] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012. 6

[13] Wayne B Nelson. *Applied life data analysis*, volume 521. John Wiley & Sons, 2003. 6

[14] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 2021. 1, 2

[15] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 2

[16] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009. 6

[17] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018. 6

[18] N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5927–5937. Curran Associates, Inc., 2017. 2, 3

[19] Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009. 6

[20] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007. 6

[21] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. A survey on optimal transport for machine learning: Theory and applications, 2021. 2

[22] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 2

[23] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013. 6

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

A. Related Works

The first step in explaining distribution shift is detecting such a shift. Many previous works have worked on this problem via methods such as statistical hypothesis testing of the input features [13, 16, 17], training a domain classifier to test between source and non-source domain samples [11], etc. However, these works' primary purpose is to provide the binary information of whether a shift has occurred or not and leave any post-detection methods up to the user (i.e., debugging and/or likely refitting a model).

In [3, 10], the authors attempt to provide more information via localizing a shift to a subset of features or causal mechanisms. [10] 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 which have shifted from source to target. This is defined using a hypothesis test which checks for any discrepancy between the conditional distributions of one feature given the rest for both the source and target distributions. The authors use $\phi_{Fisher}(P_{src}(x_j|\boldsymbol{x}_{-j}), P_{tgt}(x_j|\boldsymbol{x}_{-j}))$, as a measure of conditional divergence and report any features which have a statistically significant conditional shift from source to target. In [3], 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/discoverable. 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). How much each mechanism contributes to all possible swaps is measured (or approximated), and is deemed to be the amount that node can be "assigned blame" for the causing the change in the distribution. While both of these methods more information about distribution shift, they are mainly detection-based methods (via identifying shifted causal mechanisms or feature-level shifts), unlike an explanatory mapping which helps explain *how* the data has shifted.

The characterization of the problem of distribution shift has been extensively studied [12, 16, 19] 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 [20] the $P(\boldsymbol{x})$ marginal differs from source to target, but the output conditional remains the same, while label shift (also known as prior probability shift) [11, 23] is when the $P(y)$ marginals differ from source to target, but the full-feature conditional 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.

In contrast with current domain generalization benchmarks (e.g., WILDS [9] and DomainBed [7] benchmarks) which are focused on compiling ML train/test distribution shifts, our goal is understanding the shifts (e.g., for knowledge discovery or appropriate mitigation) rather than performing well under shifts. Thus, we even consider distribution shifts that are artificial yet interesting (like splitting the data on an attribute like gender)—or shifts based on thresholding a simulation parameter. Further, our goal likely requires shifts for which some form of ground truth explanation is known (which allows for validation of generated explanations).

B. Interpretable Transport Sets

A de facto standard practice for explaining distribution shift 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 which includes any possible mapping $T : \mathbb{R}^d \to \mathbb{R}^d$. However, even a simple transport set like Ω_{vector} can yeild uninterpretable mappings in high dimensional regimes (e.g., a shift vector of over 100 dimensions). To combat this, we can regulate the complexity of a mapping by forcing it only move points along a specified number of dimensions. We define this as 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 situations 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 which breaks the

source and target distributions into k sub-populations and provides a vector-based shift explanation per sub-population. We define this as $k-$*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 j^{th} 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 have a partial ordering on interpretability without having to determine the absolute value of interpretability of a individual explanation T, as this requires expensive context-specific human evaluations, which is out of scope for this paper.

C. Practical Methods for Finding and Validating Shift Explanations

In this section, we discuss practical methods for shift explanations. We first discuss using our k-sparse and k-cluster maps to allow a user to automatically change the level of interpretability of a shift explanation as desired. Coupled with a PercentExplained metric, this gives an operator various levels/complexities of explanation and a way to validate them. Next, we propose a practical approximation to Equation 1, the Interpretable Transport equation, and Sections C.3 and C.4 cover how to find the optimal explanation from $\Omega^{(k)}_{sparse}$ and $\Omega^{(k)}_{cluster}$ for this equation.

C.1. Interpretability as a Hyperparameter

By optimizing Equation 1 we can find the best shift explanation for a given set of interpretable transport maps Ω. However, directly defining a Ω which contains candidate mappings which are guaranteed to be both interpretable and expressive enough to explain a shift can be a difficult task. Thus, we can instead set Ω to be a super-class, such as Ω_{vector} given in Appendix B, and then 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 *PercentExplained* metric, which we define to be:

$$\text{PercentExplained}(P_{src}, P_{tgt}, T) \coloneqq \frac{W_2^2(P_{src}, P_{tgt}) - W_2^2(T_\sharp P_{src}, P_{tgt})}{W_2^2(P_{src}, P_{tgt})} \tag{3}$$

where $W_2^2(\cdot,\cdot)$ is the squared Wasserstein-2 distance between two distributions. By rearranging terms (and ignoring the percentage scaling factor) 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.

C.2. Empirical Interpretable Transport

Since the divergence term in Equation 1 can be computationally-expensive to optimize over in practice, we suggest an empirical approximation to the interpretable transport solution:

$$\arg\min_{T\in\Omega} \frac{1}{N}\sum_{i=1}^{N} c(\boldsymbol{x}^{(i)}, T(\boldsymbol{x}^{(i)}) + \lambda d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)})) \tag{4}$$

where d is a distance function such as the l_2 distance or squared euclidean distance. Most notably, the divergence value in Equation 1 is replaced with the sum over distances between $T(\boldsymbol{x})$ and the optimal transport mapping for $\boldsymbol{x}$. This is computationally attractive as the optimal transport solution only needs to be calculated once, rather than calculating the Wasserstein distance once per iteration like in the Interpretable Transport solution (which even if the W-distance is approximated, can be expensive over many iterations). For optimization purposes, this is also reasonable since $\frac{1}{N}\sum_{i=1}^{N} d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)}))$ upper-bounds $\phi(P_{T(\boldsymbol{x})}, P_{\boldsymbol{y}})$, when $d = \ell_2^2$, $\phi = W_2^2$ and N approaches the population size of P_{src} (proof shown in appendix).

C.3. Finding k-Sparse Maps

Let k be a desired level of interpretability, which for k-sparse maps is equivalent to saying $k = |\mathcal{A}(T)|$, where $\mathcal{A}$ is our active feature set (i.e., the dimensions along which our mapping can shift points). Our goal is to find the optimal k features to include in $\mathcal{A}$ and then find the best transport along those features for a given transport class Ω. A simple (and often ideal) approach to feature selection problem is to select the k features which have the largest shift in their mean from the source distribution to the target distribution; this approach is used throughout this paper. Although the chosen T will depend the optimization over Ω, we provide two closed form solutions which give optimal alignment for a given k under cases where $\Omega = \Omega_{vector}$ and when Ω is all possible mappings. The mapping which gives the best alignment in $\Omega_{vector}^{(k)}$ is k-sparse mean shift, i.e., $T(\boldsymbol{x}) = \boldsymbol{x} + \tilde{\mu}$ where $\tilde{\mu}$ is a vector where the j^{th} coordinate is $[\mu_{tgt} - \mu_{src}]_j$, if $j \in \mathcal{A}$, else, it is 0. When $\Omega^{(k)}$ is all k-sparse functions, the shift explanation which minimizes the distance term in Equation 4 is the k-sparse optimal transport solution which sets each feature in $\mathcal{A}$ to match that of the OT push forward for that feature, i.e., $[T(\boldsymbol{x})]_j = [T_{OT}(\boldsymbol{x})]_j$ if $j \in \mathcal{A}$, else $[\boldsymbol{x}]_j$. The proofs for the two previous claims can be seen in the Appendix.

C.4. Finding k-Cluster Maps

Instead of shifting respective to features, we can define k vector shifts for k groups in our source domain, with the goal of explaining how each group changed from source to target. To do this, we perform *paired* clustering in the source and target domains, so that we can relate a given cluster in P_{src} to its most similar counterpart in P_{tgt} (as opposed to pushing the k clusters in P_{src} onto the entire target domain). With this, we construct M_{src} and M_{tgt} where the k columns of M represent the k cluster means for the source and target distributions, respectively. Then, we define $\Delta = M_{tgt} - M_{src}$ so that each vector shift δ_j is the difference in means between the j^{th} source and the target clusters. In practice, the set of paired clusters can be found by performing clustering in a joint Z space of P_{src} and $P_{T_{OT}(\boldsymbol{x})}$ where the resultant k cluster centroids in this space are of the form $[M_{src}, M_{tgt}]$.

Formally, this is done using the following algorithm:

Algorithm 1 Finding k Paired Clusters

Input: X, Y, k
$d \leftarrow X.ndim$
$T_{OT} \leftarrow$ OptimalTransportAlg(X, Y) //e.g., Sinkhorn
$Z \leftarrow [X, T_{OT}(X)]$
$Z_{cluster-centroids} \leftarrow$ ClusteringAlg(Z, k) //e.g., k-means
$M_{src} \leftarrow [Z_{cluster-centroids}]_{1:d}$ //slicing column-wise
$M_{tgt} \leftarrow [Z_{cluster-centroids}]_{d:2d}$
Output: M_{src}, M_{tgt}

C.5. Proof that the distance in empirical interpretable transport upper-bounds the Wasserstein distance

First, let's remember our empirical method for finding T:

$$\arg\min_{T \in \Omega} \frac{1}{N} \sum_{i}^{N} c(\boldsymbol{x}^{(i)}, T(\boldsymbol{x}^{(i)})) + \lambda d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)})) \tag{5}$$

where T_{OT} is the optimal transport solution between our source and target domains with the given c cost function. The distance term d on the right-hand side of this equation is assumed to be the ℓ_2 cost or squared euclidean cost, and is an empirical approximation of the divergence term $\phi(P_{T(\boldsymbol{x})}, P_Y)$ in Equation 1, where ϕ is assumed to be the Wasserstein distance, W. We claim this is a reasonable approximation since as N approaches the size of the dataset (or for densities, $\lim_{N\to\infty}$),

the distance term becomes the expectation: $\mathbb{E}_{x\sim P_{src}} d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)}))$ which is an upper-bound on the $W(P_{T(\boldsymbol{x})}, P_Y)$ distance. To show this, we start with the expanded W distance:

$$
\begin{aligned}
W(P_{T(\boldsymbol{x})}, P_Y) &= \min_{R\in\Psi} \mathbb{E}_{\boldsymbol{x}\sim P_{src}} d\left(T(\boldsymbol{x}), R(T(\boldsymbol{x}))\right), \quad \Psi \coloneqq \{R : R_\sharp T(\boldsymbol{x}) = P_Y\} \\
&\leq \mathbb{E}_{\boldsymbol{x}\sim P_{src}} d\left(T(\boldsymbol{x}), R(T(\boldsymbol{x}))\right), \qquad \forall R \in \Psi \\
&\text{If we let } Q = T_{OT}\cdot T^{-1}, \text{ and since } Q \in \Psi \text{ we can say} \\
&\leq \mathbb{E}_{\boldsymbol{x}\sim P_{src}} d\left(T(\boldsymbol{x}), Q(T(\boldsymbol{x}))\right) = \mathbb{E}_{\boldsymbol{x}\sim P_{src}} d\left(T(\boldsymbol{x}), T_{OT}(\boldsymbol{x})\right) \\
\implies W(P_{T(\boldsymbol{x})}, P_Y) &\leq \mathbb{E}_{\boldsymbol{x}\sim P_{src}} d\left(T(\boldsymbol{x}), T_{OT}(\boldsymbol{x})\right)
\end{aligned}
$$

C.6. Proving the k-sparse optimal transport is the k-sparse transport that minimizes our distance from OT loss

When performing unrestricted k-sparse transport, i.e., where $\Omega^{(k)}_{sparse}$ is any transport which only moves points along k dimensions, the k-sparse optimal transport solution is the exact mapping that minimizes the distance function in the right-hand side of Equation 5 if d is the ℓ_2 distance or squared Euclidean distance. As a reminder, k-sparse optimal transport is: $[T(\boldsymbol{x})]_j = [T_{OT}(\boldsymbol{x})]_j$ if $j \in \mathcal{A}$, else $[\boldsymbol{x}]_j$, where $\mathcal{A}$ is the active set of k dimensions which our k-sparse transport T can move points. Let $\bar{\mathcal{A}}$ be $\mathcal{A}$'s compliment (i.e. the dimensions which are unchanged under T). 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:

$$
\begin{aligned}
d(\boldsymbol{z}, \boldsymbol{z}_{OT}) &= \sum_{j\in[d]}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 + \underbrace{\sum_{j\in\bar{\mathcal{A}}}\sum_{i\in[n]} \left(\boldsymbol{x}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2}_{=\alpha \text{ , since constant w.r.t T}} \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 + \alpha \\
&\text{now if T is the truncated optimal transport solution, } [\boldsymbol{z}]_j = [\boldsymbol{z}_{OT}]_j \quad \forall j \in \mathcal{A} \\
&= 0 + \alpha
\end{aligned}
$$

Since α is the minimum of $d(\boldsymbol{z} - \boldsymbol{z}_{OT})$ for a given $\mathcal{A}$, the truncated optimal transport problem minimizes the $d(T(\boldsymbol{x}^{(i)}), T_{OT}(\boldsymbol{x}^{(i)}))$ distance. This can easily be extended to show that the optimal active set for this case is the one that minimizes α, thus the active set should be the k dimensions which have the largest squared difference between $\boldsymbol{x}$ and $\boldsymbol{z}_{OT}$.

C.7. Proof that k-mean shift is the k-vector shift that yields the best alignment

When performing k-sparse vector transport, i.e., where $\Omega^{(k)}_{vector} = \{T : T(\boldsymbol{x}) = \boldsymbol{x} + \tilde{\delta}\}$ where $\tilde{\delta} = [\delta]_j$ if $j \in \mathcal{A}$ else $[\delta]_j = 0$ and $\delta \in \mathbb{R}^d$, $|\mathcal{A}| \leq k$, the k-sparse mean shift solution is the exact mapping that minimizes the distance function in the right-hand side of Equation 5 when d is the ℓ_1 distance.

$$
\begin{aligned}
d(\boldsymbol{z}, \boldsymbol{z}_{OT}) &= \sum_{j\in[d]}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 + \underbrace{\sum_{j\in\bar{\mathcal{A}}}\sum_{i\in[n]} \left(\boldsymbol{x}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2}_{=\alpha \text{ , since constant w.r.t T}} \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{z}_{i,j} - \boldsymbol{z}_{OT_{i,j}}\right)^2 + \alpha \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{x}_{i,j} + \delta_j - \boldsymbol{z}_{OT_{i,j}}\right)^2 + \alpha \\
&= \sum_{j\in\mathcal{A}}\sum_{i\in[n]} \left(\boldsymbol{x}_{i,j}^2 + \delta_j^2 + \boldsymbol{z}_{OT_{i,j}}^2 + 2\delta_j(\boldsymbol{x}_{i,j} - \boldsymbol{z}_{OT_{i,j}}) - 2\boldsymbol{z}_{OT_{i,j}}\delta_j - 2\boldsymbol{x}_{i,j}\boldsymbol{z}_{OT_{i,j}}\right) + \alpha
\end{aligned}
$$

Similar to the k-sparse optimal transport solution, we can see that $\mathcal{A}$ should be selected as the k dimensions which have the largest shift, thus minimizing α. The coordinate-wise gradient of the above equation is:

$$
\nabla_{\delta_j}\, d(\boldsymbol{z}, \boldsymbol{z}_{OT}) = \begin{cases} \sum_{i\in[n]} \left(2\delta_j + 2\boldsymbol{x}_{i,j} - 2\boldsymbol{z}_{OT_{i,j}}\right) & j \in \mathcal{A} \\ 0 & j \in \bar{\mathcal{A}} \end{cases}
$$

Now with this we can say:

$$
\begin{aligned}
\nabla_{\delta_{j\in\mathcal{A}}}\, d(\boldsymbol{z}, \boldsymbol{z}_{OT}) &= \sum_{i\in[n]} \left(2\delta_j + 2\boldsymbol{x}_{i,j} - 2\boldsymbol{z}_{OT_{i,j}}\right) \\
&= 2n\delta_j + \sum_{i\in[n]} \left(2\boldsymbol{x}_{i,j} - 2\boldsymbol{z}_{OT_{i,j}}\right) \\
&\text{now let } \delta_j = \delta_j^* \\
0 &= 2n\delta_j^* + \sum_{i\in[n]} \left(2\boldsymbol{x}_{i,j} - 2\boldsymbol{z}_{OT_{i,j}}\right) \\
n\delta_j^* &= \sum_{i\in[n]} \left(\boldsymbol{z}_{OT_{i,j}} - \boldsymbol{x}_{i,j}\right) \\
\delta_j^* &= \frac{1}{n}\sum_{i\in[n]} \left(\boldsymbol{z}_{OT_{i,j}} - \boldsymbol{x}_{i,j}\right) \\
\delta_j^* &= \mu_{\boldsymbol{z}_{OT_j}} - \mu_{\boldsymbol{x}_j}
\end{aligned}
$$

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

D. Additional Experiment Details and Results

Here we provide more raw samples from the ColorMNIST experiment as well as an additional counterfactual example experiment, but this time on a toy dataset (as opposed to the real world experiment seen in subsection 3.2) to illustrate the power of distributional counterfactual examples.

D.1. Additional Counterfactual Example Experiment to Explain a Multi-MNIST shift

As mentioned in subsection 3.2, 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 3.2) 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 [6] 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.

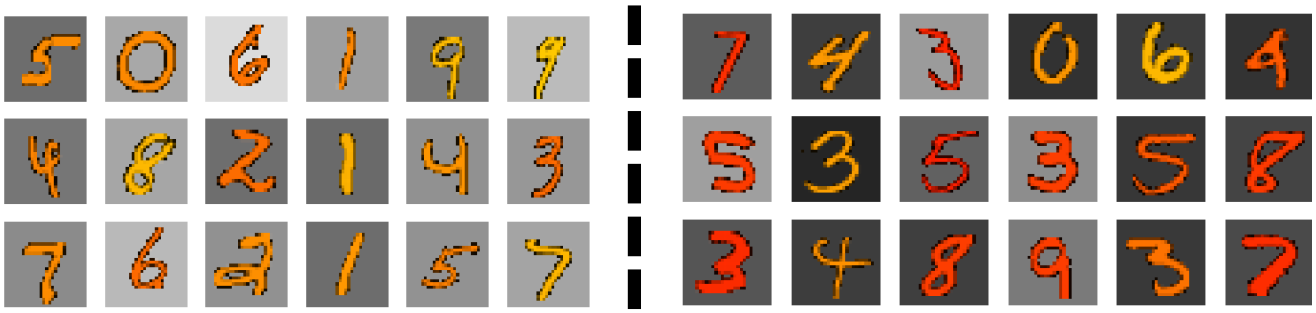

Figure 3. Samples from the source environment (left) with more yellow digits and lighter backgrounds while the target environment (right) has more red digits and/or darker backgrounds.

To generate the counterfactual examples, we use a Domain Invariant Variational Autoencoder (DIVA) [8], 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 4 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).

Algorithm 2 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$(\boldsymbol{x}_2)$
$\hat{\boldsymbol{x}}_{1\to 2} \leftarrow$ model.decode$(z_{y_1}, z_{d_2}, z_{residual_1})$
$\hat{\boldsymbol{x}}_{2\to 1} \leftarrow$ model.decode$(z_{y_2}, z_{d_1}, z_{residual_2})$
Output: $\hat{\boldsymbol{x}}_{1\to 2}$, $\hat{\boldsymbol{x}}_{2\to 1}$

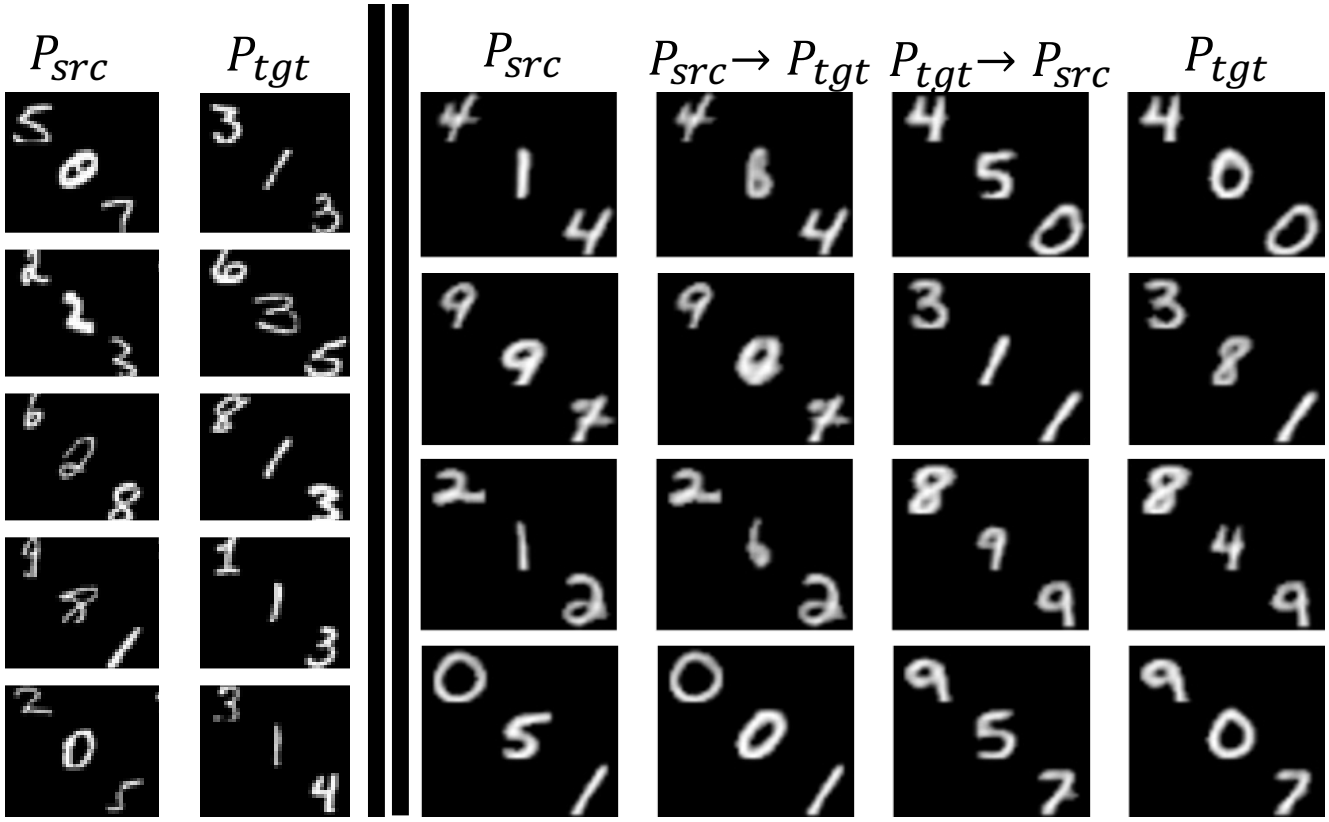

Figure 4. A comparison of the baseline grid of unpaired source and target samples (left) and counterfactual pairs (right) which shows how counterfactual examples can highlight the difference between the two distributions. For each image, the top left digit represents the class label, the middle digit 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.