Out of Distribution Generalization via Interventional Style Transfer in Single-Cell Microscopy

Wolfgang M. Pernice*
Columbia University
wp2181@cumc.columbia.edu

Michael Doron
Broad Institute
mdoron@broadinstitute.org

Alex Quach
MIT
aquach@mit.edu

Aditya Pratapa
Broad Institute
adyprat@pm.me

Sultan Kenjeyev
University College London
ksula0155@gmail.com

Nicholas De Veaux
New York University
nrdeveaux@gmail.com

Michio Hirano
Columbia University
mh29@cumc.columbia.edu

Juan C. Caicedo*
Broad Institute
jcaicedo@broadinstitute.org

Abstract

Real-world deployment of computer vision systems, including in the discovery processes of biomedical research, requires causal representations that are invariant to contextual nuisances and generalize to new data. Leveraging the internal replicate structure of two novel single-cell fluorescent microscopy datasets, we propose generally applicable tests to assess the extent to which models learn causal representations across increasingly challenging levels of OOD-generalization. We show that despite seemingly strong performance as assessed by other established metrics, both naive and contemporary baselines designed to ward against confounding, collapse to random on these tests. We introduce a new method, Interventional Style Transfer (IST), that substantially improves OOD generalization by generating interventional training distributions in which spurious correlations between biological causes and nuisances are mitigated. We publish our code[1] and datasets[2].

1. Introduction

The ability to learn meaningful visual features from multiplexed microscopy images of cells and tissues promises to unlock cellular morphology as a powerful new single-cell omic with considerable potential to advance biomedical research [28]. In turn, efforts are underway to collect fluorescent microscopy datasets that interrogate single-cell biology across hundreds of millions of cells and thousands of biological perturbations [3, 8]. To enable scientific discovery, computer vision models must learn representations that generalize to observations made in new observational environments (OEs) [33, 34]. Yet, vision systems are prone to learning spurious correlations between concepts of interest (e.g. objects) and contextual nuisances (e.g. background) [24]. This can yield biased representations that, although they may generalize well to hold-out sets that are independent and identically distributed (IID) with respect to the training data, collapse when tested on data that fall outside this distribution. For example, the performance of state-of-the-art (SOTA) vision models trained on the stereotypical views of objects in ImageNet, dramatically deteriorates when tested on ObjectNet images [1], which were collected with proactive interventions on several nuisance factors, such as background and object orientation (e.g. fallen-over chairs), that pose little challenge to humans.

The same confounding influence that OEs exhibit in natural image datasets, manifests in biomedical datasets in the form of "batch-effects". Indeed, despite best efforts, technical variation between datasets collected in separate (experimental) batches cannot be perfectly controlled. Given the susceptibility of vision models to spurious correlations in natural image data outlined above, batch-effects present a major threat to meaningful biomedical applications of representation learning in fluorescent microscopy.

In this paper we adopt the terminology of causal inference [29, 33] to study the (batch)effects of OEs as a confounder $\mathbf{C}$ to a general causal process that we suggest describes most datasets in biology. Our goal is to learn representations from such data, that model causal relationships,

*Co-corresponding authors
[1] https://github.com/Laboratory-for-Digital-Biology/IST
[2] 10.5281/zenodo.7830240

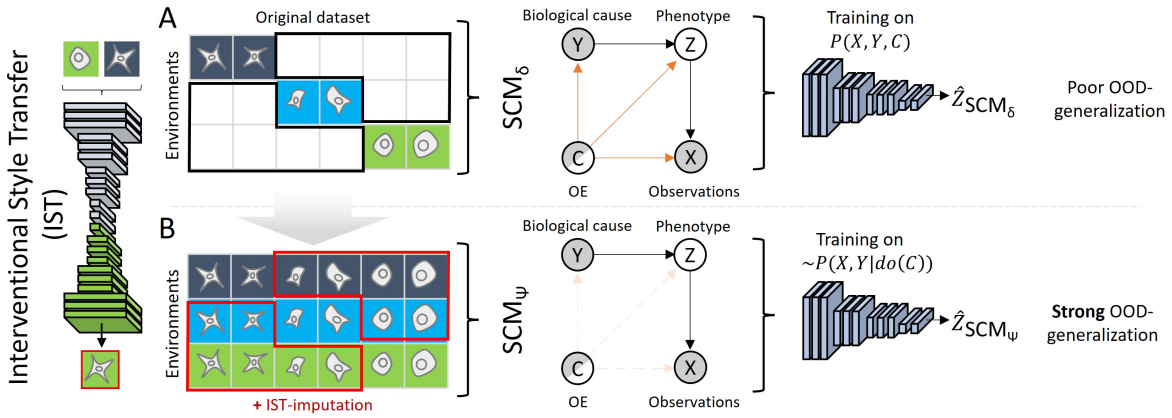

Figure 1. (A) In most high-content datasets, not all conditions are observed in all OEs. Training distributions thus entail spurious correlations between OEs, observations and biological causes, yielding models that learn confounded representations according to $\mathbf{SCM}_\delta$. (B) We introduce an IST approach to impute images as if they had been collected in different OEs. By randomly permuting images across OEs, we yield an interventional distribution that removes spurious correlations with OEs allowing models to learn representations that are less biased and better capture the true causal structure according to $\mathbf{SCM}_\psi$.

while remaining invariant to OEs. Importantly, we hold that the hypothesis that a given representation is causal (i.e. invariant over OEs and nuisances) cannot be falsified using IID hold-out data. Instead, we propose a rigorous testing regime based on generalization to OEs which are out-of-distribution (OOD) compared to the training data as a necessary characteristic and critical empirical measure of causal learning. To this end, and to foster progress towards causal representation learning in the field, we publicly release two real-world single-cell fluorescent microscopy datasets that exhibit internal replicate structures representative of most high-content imaging protocols (see below). We leverage this substructure to design realistic OOD generalization tasks. Surprisingly, we find that not only naive, but also SOTA post-processing and regularization baselines designed to mitigate batch-effects and improve generalization, fail when evaluated on these OOD tasks, despite in part excellent scores on IID hold-out sets and auxiliary metrics.

Given the ineffectiveness of existing methods on our OOD-task, we next consider intervening on the training distribution itself. Intuitively, if the training set contained observations balanced over all OEs, models should learn invariances to OEs and represent the right causal structure [24, 29]. While collecting such dense datasets is not impossible (see e.g. [32]), many key applications require the assessment of very large numbers of conditions (e.g. over even modestly-sized drug libraries) that, for practical reasons, have to be collected in multiple batches. As such, most high-content imaging datasets are sparse, that is, sets of conditions are only observed in some OEs but not others (see Figs. 1, 3). Inspired by recent results on generative interventions to mitigate biases in natural image data [24], we propose a new, light-weight method for *Interventional Style Transfer* (IST) that generates effective interventions across an arbitrary number of OEs. To achieve this, we introduce architectural innovations and loss terms that prevent content hallucinations, which we find leads to failure of other style-transfer methods on our benchmark datasets. We then employ IST to yield a training distribution that mitigates OEs as confounders (Fig.1) and show that models trained on it exhibit major improvements in OOD-generalization.

As our main contributions, we (1) publish two new benchmark single-cell datasets with different degrees of sparsity in their replicate structure; we (2) propose a rigorous OOD-generalization test regime that can be adopted across most experimental dataset; and (3) we introduce IST as the first method that achieves substantial improvements across increasingly challenging levels of OOD-generalization, as a starting-point for future work towards causal representation learning in microscopy and beyond.

2. Related Work

Data Augmentation: Data augmentations, such as blur, contrast, and rotations, are almost universally used in computer vision to yield more robust models [21, 35]. Both style-transfer [10, 36] and adversarial training [38] have been employed in the pursuit of more complex augmentations. Our IST approach can be viewed as learning augmentations that imitate the effect of confounders.

Generative Models: Generative models have been successfully employed on fluorescent microscopy and other biomedical data [12, 39]. When OEs are unobserved, [24] show that generative models can be steered to produce noisy image manipulations on complex nuisances such as view point, that, when employed during training, improve OOD generalization. We employ the known replicate structure of our data to steer the generator directly.

Domain Adaptation: Our approach builds on advances in domain-adaptation and style-transfer, developed to allow for the differential manipulating a style while preserving other content [4, 5, 18, 20]. A major risk in applying style-transfer methods to scientific data is the inadvertent alteration of content (in our case phenotypic information). We hence design our IST approach to emphasizes content-preservation by discouraging major changes in pixel space.

Batch-effect correction: How to mitigate batch-effects is an active field of study in biomedicine [17]. Our work is closest to [30] who employ style-transfer to disentangle batch effects from biological features. IST features architectural improvements that prevent content alterations without the need for threshold- or segmentation-based regularization terms. IST also does not depend on assumptions about the nature of batch-effects (such as that they primarily manifest in first-order statistics [7]), and achieves strong performance on challenging benchmarks without the need for additional post-hoc methods employed in [30].

Fairness: A considerable body of work in visual recognition explores questions of fairness e.g. over demographic factors [11, 40], including by means of style-transfer. Although discussions of causality are absent from these works, questions of fairness relate to our study on batch-effects as we seek to learn causal representations from biased data.

3. Causal Analysis

Fig. 1 presents a generalized structural causal model (SCM) [29] that we assume as the basis of our work. We seek to reveal causal relationships between a set of conditions $\mathbf{Y}$ (e.g. disease categories) that may manifest cellular phenotypes $\mathbf{Z}$. To characterize $\mathbf{Z}$, we collect observations $\mathbf{X}$ using fluorescent microscopy. Observations are made in specific OEs $\mathbf{C}$ (i.e. batch, constituted by a specific well, plate, aliquot of reagents, etc.) that introduce technical variation to $\mathbf{X}$, and may further influence the biology of $\mathbf{Z}$, revealing it as a confounder [33]. Importantly, in most datasets, not all conditions (we say, biological causes) are observed in all OEs. As such, the specific OE $\mathbf{c}$ also determines (the set of) biological causes $\mathbf{Y}$: e.g. two plates may contain different sets of conditions. Given training distributions $P(X, Y, C)$ in which biological causes are sparse over OEs, discriminative models learn spurious correlations between biological causes $\mathbf{Y}$ and confounding OEs $\mathbf{C}$ according to $\mathbf{SCM}_\delta$ resulting in biased representations $\hat{\mathbf{Z}}_{\mathbf{SCM}_\delta}$ that generalize poorly to new OEs (Fig. 1A).

A method that could produce faithful imputations of source images from one OE as if they had been collected in a different OE, could allow us to approximate the interventional distribution $P(X, Y|do(C))$ that would eliminate the backdoor paths emanating from $\mathbf{C}$, removing OEs as a confounder [29]. Recent work on generative interventions for causal learning in natural images showed that even under noisy image manipulations, a classifier can learn better features for recognition in OOD data [24]. Inspired by their results, we propose an interventions-based approach that is compatible with experimental datasets. Importantly, in most natural image data-generation processes, observations cause labels, i.e. human experts label images according to what they see, and we train models to recapitulate this ability [24, 31]). Instead, in our datasets, $\mathbf{Y}$ represents conditions that we *hypothesize* may cause observable cellular phenotypes. Our goal is then to approximate the conditional distribution $P(Y|X)$, that is to estimate the cause $\mathbf{Y}$ given noisy observations $\mathbf{X}$, whereby we hope to learn (discover) biologically meaningful representations of a priori unknown phenotypes $\mathbf{Z}$. Second, in contrast to natural images, where OEs are generally unobserved, our experimental data-acquisition protocols inherently document a rich ontology of processing steps that lead to any particular image (Fig. 3A). OEs are thus systematically tracked and feature rich metadata through which $\mathbf{C}$ is partially observed. In contrast to [24], we can hence explicitly steer the data generation process learned by IST according to the known OE structure of our datasets. By using IST to intervene on $\mathbf{C}$, we seek to mitigate spurious correlations in the training distribution, yielding $\hat{\mathbf{Z}}_{\mathbf{SCM}_\psi}$ and representations that generalize to OOD data (Fig. 1B).

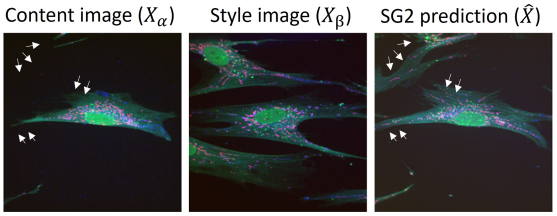

Figure 2. StarGANv2 produces compelling images, but introduces both subtle and more obvious content alterations (white arrows)

4. Method

Advances in neural style transfer make it possible to perform image transformations that preserve spatial content while adjusting other feature statistics as desired [4, 5, 18, 22]. In order to generate effective interventions on $\mathbf{C}$, a style-transfer model must learn to specifically transfer features related to OEs, while preserving phenotypic content. StarGANv2 [5] introduced a framework to train a single encoder-decoder architecture capable of style-transfer between an arbitrary number of style-domains, such as demographic categories. We instead aim to steer our model to generate images in the "style" of specific OEs. Indeed, we find that StarGANv2, adapted to multichannel microscopy, produces visually compelling images. However, the outputs consistently feature both subtle and more obvious content hallucinations (Fig. 2), suggesting that StarGANv2 fails to

adequately preserve phenotypic content.

As a potential alternative to style-transfer in pixel space, [7] propose MixStyle to pursue domain-generalization by mixing style-features of images from different OEs in the feature-maps of hidden-layers during training, with promising results. Remarkably, this avoids the need for a generator all together, making Mixstyle extremely lightweight. However [7] assume that computing mean and standard-deviation of the feature-maps is sufficient to adequately capture style. While this may hold to some extent for natural images [15], there are no guarantees that this is true for batch-effects in microscopy data, and indeed, we find MixStyle offers little-to-no benefit on our task (see Sec. 6).

To avoid these failure modes we design IST with several major and minor innovations. Specifically, to enforce content preservation, we introduce skip-connections between bottle-neck layers of our encoder and decoder, and introduce three complementary loss terms that discourage phenotypic alterations. Second, we find that, although first-order feature-map statistics are by themselves insufficient to describe OEs, they suffice as style-codes that, when injected into Adaptive Instance Normalization (AdaIN) layers [15], can be interpreted by our decoder to generate output images across an arbitrary number of OEs 4. This allows us to avoid all auxilliary networks required by StarGANv2 or comparable methods, rendering IST not only effective, but also computationally efficient, as detailed below.

4.1. Model Components

To train our IST-model, let $\mathbf{X}$ be the set of images, with associated environments $\mathbf{C}$ and cause labels $\mathbf{Y}$, respectively. Given an image $x \in \mathbf{X}$ observed in environment $c \in \mathbf{C}$, we seek to train a Generator $\mathcal{G}$ capable of producing image transformations $\hat{x}$ as if they come from other environments $\mathbf{C}$ (style) while preserving the original phenotypic content $z \in \mathbf{Z}$. To this end, we derive *style codes* v and train $\mathcal{G}$ to interpret them. Our framework consists of three main modules (see Fig. 4A):

Encoder: Given an image x, the encoder $\mathcal{E}$ derives the representation $u_x = \mathcal{E}(x)$, composed of multi-layer feature-sets u_x^l, with $l \in \{1..., L\}$ and L the number of layers in the network. We implement $\mathcal{E}$ as a ResNet18 [14] with instance normalization (IN) layers and a few modifications to facilitate skip connections. We pre-train $\mathcal{E}$ on an auxiliary multi-task objective of predicting C and Y given X.

Generator: Our generator predicts output images $\hat{x} = \mathcal{G}(u, v)$ given a feature set u and an style code v. To promote the preservation of phenotypic content in the output images, we bias $\mathcal{G}$ against major changes in pixel space [16], by implementing $\mathcal{G}$ as a UNet-decoder with skip connections that concatenate feature-sets u_x^l from the l-th corresponding feature-layer of the encoder, with $l \in \{1..., L\}$. We find that our choice improves both similarity between pairs x and $\hat{x}$ as well as the realism of our output images.

Critic: Similar to [5] we implement a critic $\mathcal{D}$ as a multi-task discriminator with N_c output heads, where N_c is the number of OEs. Each head $\mathcal{D}_c$ is trained as a binary classifier to distinguish real from fake images of their true or assigned OE c. To facilitate convergence, we initialize $\mathcal{D}$ with the weights of the pre-trained encoder $\mathcal{E}$ and fine-tune over the adversarial optimization process.

Style codes: To steer our generator, we compute image-specific style-codes. StarGANv2 employs a dedicated style-encoder to derive style-codes from latent distributions or input images [5]. We find that effective style-codes can be computed directly from image features using our pre-trained, frozen encoder $\mathcal{E}$. Given the features $u_i = \mathcal{E}(x_i)$ of an image x_i, we compute:

$$v_i = \left[(\mu_{u_i^l}, \sigma_{u_i^l}) : l \in \{1...L\} \right] \tag{1}$$

where $\mu_{u_i^l}$ and $\sigma_{u_i^l}$ are the mean and standard deviation across the spatial domain of the feature maps of layer l.

4.2. Training the Generator

We train $\mathcal{G}$ to transform the appearance of an image from one OE to another. Following pre-training, we freeze the encoder $\mathcal{E}$ and use its weights to initialize the critic $\mathcal{D}$. The generator $\mathcal{G}$ is trained using SGD on pairs of triplets: $(x_\alpha, y_\alpha, c_\alpha)$ for content images and $(x_\beta, y_\beta, c_\beta)$ for style images. During training, we randomly sample content images balanced over $\mathbf{Y}$, and style images balanced over $\mathbf{C}$. In that way, content and style images are independently drawn to ensure samples with diverse phenotypic and technical variation respectively. During the forward pass, we first compute their feature sets $u_\alpha = \mathcal{E}(x_\alpha)$ and $u_\beta = \mathcal{E}(x_\beta)$ to then derive style codes v_α, v_β. To intervene on the OE of the content image, we then inject v_β using AdaIN-layers to predict the output $\hat{x} = \mathcal{G}(\mathcal{E}(x_\alpha), v_\beta)$. We minimize the following training objectives:

Adversarial Loss: Given a pair of content and style images x_α, x_β, we compute style-codes v_α, v_β as described above. The generator is trained to produce realistic output images $\hat{x} = \mathcal{G}(\mathcal{E}(x_\alpha), v_\beta)$ with the following adversarial loss:

$$\begin{aligned} \mathcal{L}_{Adv} =& \mathbb{E}_{x,c} \left[\log(\mathcal{D}_c(x)) \right] + \\ & \mathbb{E}_{x,\tilde{x},\tilde{c}} \left[\log(1 - \mathcal{D}_{\tilde{c}}(\mathcal{G}(\mathcal{E}(x), v_{\tilde{c}}))) \right], \end{aligned} \tag{2}$$

where $\mathcal{D}_c(\cdot)$ is the head corresponding to OE c. $\mathcal{G}$ learns to use the style-codes $v_{\tilde{c}}$ to generate versions of x as if observed in another environment $\tilde{c}$.

Style Loss: We further ensure effective intervention by applying a style-loss:

$$\mathcal{L}_{Style} = \mathbb{E}_{\hat{x},c} \left[\frac{1}{L} \sum_{l=1}^{L} \| \text{Gram}(u_x^l) - \text{Gram}(u_{\hat{x}}^l) \|_1 \right] \tag{3}$$

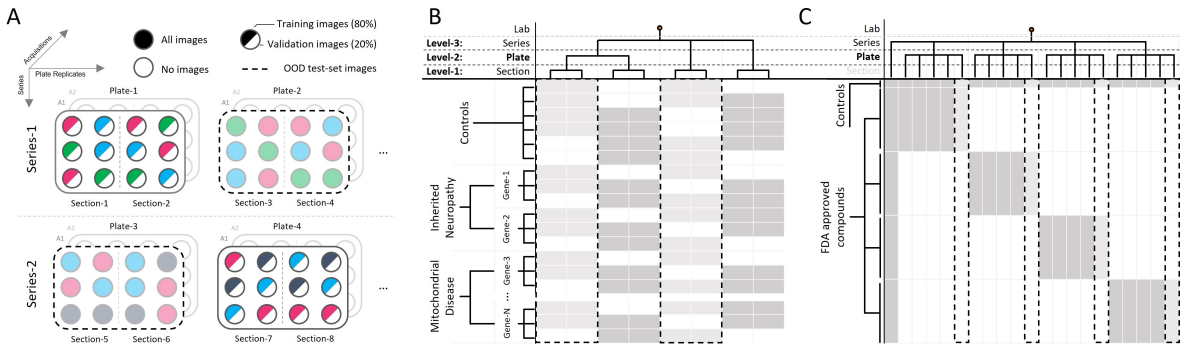

Figure 3. A: diagrammatic illustration of a generalized data-acquisition process for high-content microscopy. Well colors indicate conditions (e.g. cell lines or perturbations) arrayed over multiwell plates with limited capacity. Datasets exhibit a nested replicate structures; a series constitutes a full experimental replicate including fresh cells and reagents, which may contain further replicates by plate, plate-section (acquired separately), and acquisitions, each constituting potentially meaningful OEs. This yields datasets of varying degrees of sparsity. B,C: schematic dataset substructure for GRID and LINCS-SC respectively. We define *levels of generalization* according to increasingly distant relationships between OEs. The training and OOD-test setup for one fold of level-2 is indicated.

where Gram$(\cdot)$ denotes the Gram matrix of features in the l-th layer of the encoder $\mathcal{E}$, used in style transfer to match the feature covariance of stylized images [22, 23].

Cycle-Consistency Loss: To promote the preservation of phenotypic content of a source image x_α in the output $\hat{x} = \mathcal{G}(\mathcal{E}(x_\alpha), v_\beta)$, we apply a cycle-consistency loss [41]:

$$\mathcal{L}_{Cyc} = \mathbb{E}_{x,c}[\|x - \mathcal{G}(\mathcal{E}(\hat{x}), v_c)\|_1], \tag{4}$$

where v_c is the estimated style code of the original content image, i.e. we reconstruct x_α from $\hat{x}$.

Content Loss: We additionally constrain the absolute changes in pixel space between x and $\hat{x}$ to prevent substantial loss or addition of phenotypic content (e.g. the hallucination of new cells or cellular components) by applying a content loss

$$\mathcal{L}_{Cont} = \mathbb{E}_{\tilde{x},c}\left[\|(\hat{x} - x)\|_1 + \frac{1}{L}\sum_{l=1}^{L}\|z_{\hat{x}}^l - z_x^l\|_1\right] \tag{5}$$

Class-matching Loss: To further enforce that the generator preserves the phenotypic characteristics of input images, we implement a class-matching loss, defined as:

$$\mathcal{L}_{Cmatch} = \mathbb{E}_{\tilde{x}}\left[-\sum_{y \in Y} \hat{y}\, log\left(\mathcal{E}_{cmt}(\hat{x})_y\right)\right], \tag{6}$$

which is essentially the cross entropy loss of the cause predictions for the synthesized image with respect to the predictions for the real input image, according to the frozen encoder classifier $\mathcal{E}_{cls}$. Note that instead of using the actual cause label y, we use as target the prediction for the real image $\hat{y} = \mathcal{E}_{cls}(x_c)_y$.

Full Objective: We then optimize a min-max objective that trains the generator and critic in an adversarial fashion:

$$\begin{aligned}\min_{\mathcal{G}} \max_{\mathcal{D}} =& \mathcal{L}_{Adv} + \lambda_1 \mathcal{L}_{Style} + \\ & \lambda_2 \mathcal{L}_{Cyc} + \lambda_3 \mathcal{L}_{Cont} + \lambda_4 \mathcal{L}_{Cmatch},\end{aligned} \tag{7}$$

where $\lambda_i \in \mathbb{R}$ are hyperparameters of the loss terms.

4.3. IST for causal learning

Once our IST-model is trained, we employ it to generate an interventional training distribution $P(\hat{X}, Y|do(C))$, on which we in turn train a predictor network P (Fig. 4A). To produce $\hat{X}$ during predictor training, content x_α and style x_β images are sampled from the training distribution by pairing random causes with random OEs (both drawn uniformly) and passed through the (frozen) IST-model. This strategy breaks the spurious correlations between biological causes and OEs present in the original datasets. During testing, we also pass test images through our IST-model by randomly pairing them with a training image. This can be interpreted as bringing unseen images to familiar OEs for analysis, and we observed that IST-trained predictors perform better using this additional test-time correction.

5. Experiments

To evaluate the merits of our IST approach in causal representation learning compared to relevant contemporary baselines, we conduct experiments on two novel single-cell microscopy datasets that exhibit different degrees of sparsity and correlation between biological causes and OEs (Fig. 3B,D). Based on the known OE substructure in these dataset, we propose and empirically assess three increasingly challenging levels of OOD-generalization, by constructing hold-out sets according to a hierarchy of process-

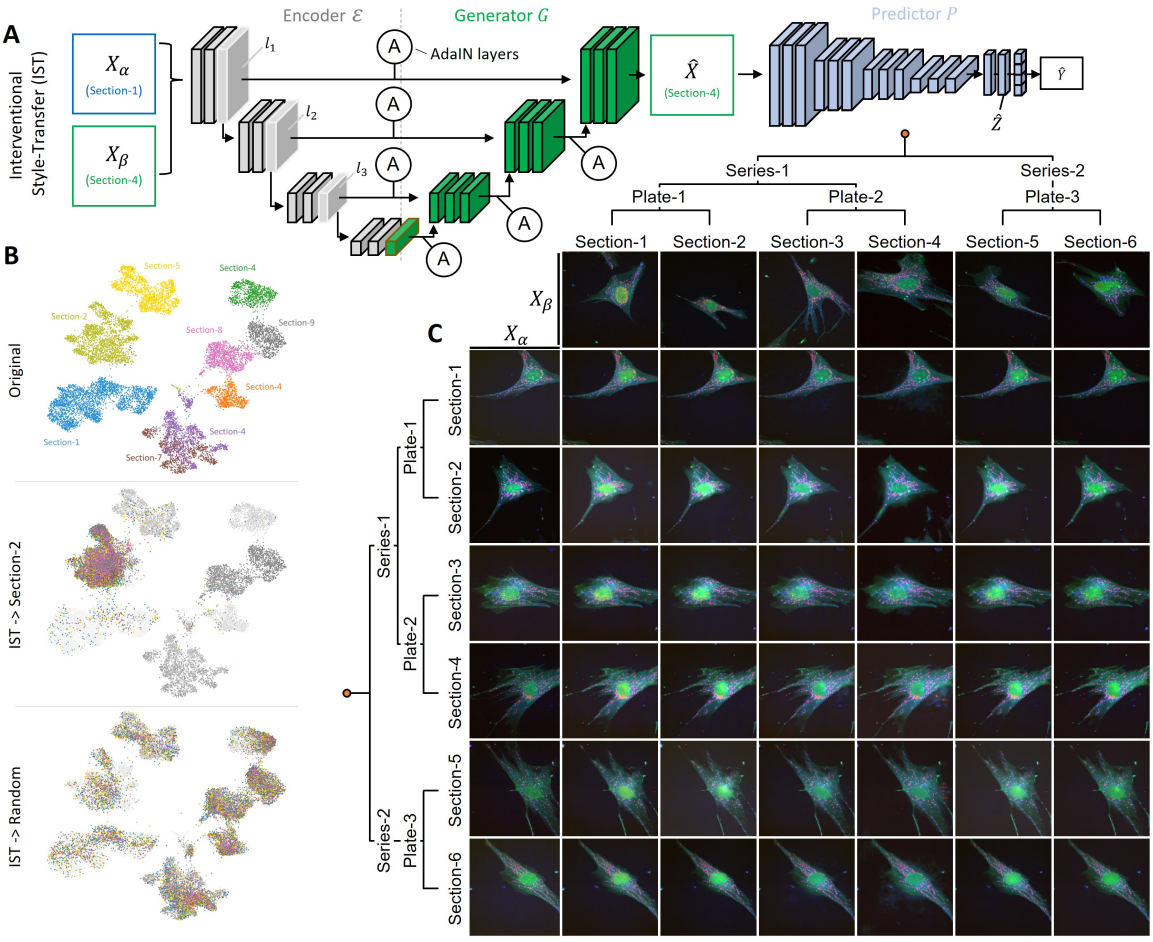

Figure 4. A: Diagram of our IST-method. Given images x_α and x_β, encoder $\mathcal{E}$ (gray) extracts latent representations u_α and provides it to our generator (green) G. $\mathcal{E}$ further extracts style-codes x_β and provides them to G via AdaIN layers. We yield a prediction $\hat{x}$ that preserves the phenotypic content of x_α but inherits the OE of x_β. We train predictors P (blue) on the resulting data distribution $\hat{X}$. B, C: UMAPs and output images illustrating the capacity of IST to project images into specific batches. When x_β is sampled from a specific OE, output images fall onto their expected landmark in the UMAP space computed on the pretrained representations of $\mathcal{E}$ (see Sec. 4). When sampling x_β fairly from all training OEs, the resulting distribution $\hat{X}$ is randomized over all OEs.

ing steps that separate them from the training data (Fig.3).[3]

5.1. Datasets

GRID: We publish a subset of the *Genetics of Rare Inherited Disease* (GRID) dataset, collected to discover latent disease-associated phenotypes in primary patient cells. The dataset contains 17,030 fluorescent microscopy images that reveal the organelle structure of primary dermal fibroblasts derived from 19 patients with 8 genetically confirmed inherited mitochondrial or neuromuscular diseases, and healthy controls (Fig.3B). Data was acquired in multi-well plates with a hierarchical replicate structure: images were collected within the minimal OE of individual wells that contain cells of a specific cell-line. Images of the same cell-lines were collected in multiple (replicate) wells, organized into plate-*sections*, *plates*, and *series* (Fig. 3A,B). Replicate wells across *sections* (level-1) are seeded onto the same plate, during the same tissue culture session and derive from the same source cultures. Plate-level replicates (level-2) are separated by *plate*, but share source cultures. Finally, *series* (level-3) indicate full experimental replicates, starting with fresh thaws of cells. Critically, while sections contain identical sets of cell-lines, they only partially overlap between plates and series, yielding a sparse matrix of biological causes vs. OEs (Fig. 3B).

LINCS-SC: In contrast to GRID, the LINCS Cell-Profiling dataset was collected as a pharmacological perturbation study, including 1,327 clinically relevant compounds [6], using a *single* A549 lung cancer cell line [37]. Cells were stained according to the Cell-Painting protocol [2] and im-

[3] Although we offer some details on biological causes for context, we do not interpret our results with respect to their biological implications; we focus on testing the generality of models across OEs.

aged at lower magnification, such that the resulting images contain many cells. The LINCS single-cell (LINCS-SC) dataset, contains a subset of 101 compounds with strong morphological effects as judged by prior analyses [27]. Single-cell images were derived by segmenting source images with Cell Profiler [26] for a total of 200,000 images. In contrast to GRID, LINCS plates contain no sections. Moreover, LINCS plate- and series-level replicates are structured according to 25 unique plate-maps that host exclusive perturbations: with the exception of controls, there is no overlap between compounds across plate-maps. Consequently, the data-matrix is almost perfectly sparse between plate-maps (Fig. 3C). Finally, in LINCS, only one series contains all plate-maps (i.e. treatments), but without plate-level replicates, while four additional series contain exclusive subsets of plate-maps, each replicated 5 times (Fig. 3C).

5.2. Baselines

We seek to train predictors P such that they generalizes to unseen OEs. We compare IST to strong domain-specific and more general baselines that collectively represent three major categories: post-hoc correction in feature space, regularization during training, and interventions on the training distribution. For all experiments, we use the same set of pre-processing steps and augmentations. As a naive baseline, we randomly initialize a predictor P as a ResNet18 network (using IN layers) attached to a linear classification head and train it to predict biological causes $\mathbf{Y}$ from $\hat{\mathbf{X}}$. We implement other methods via minimal necessary deviations:

Symphony: Symphony (SYM) is a state-of-the-art batch-effect correction method developed for single-cell RNA-sequencing (scRNAseq) datasets [17]. Symphony extends a previous method, Harmony [19], which learns linear corrections over labeled nuisances. In contrast to Harmony, Symphony allows for inference on unseen datasets. We fit Symphony on training-set features $\hat{Z}$ extracted from our naive baseline. We set the *topn* hyperparameter equal to our feature-dimension and empirically tune others.

Domain-Adversarial Regularization: We also compare to domain-adversarial (DR) training as a a regularization technique to learn features that discriminate classes but are invariant to domain-shifts between datasets [9]. We adopt this strategy with slight modification to allow for multiple domains (OEs). Specifically, we modify the architecture of our naive baseline by adding a second classification-head that distinguishes OEs. During backpropagation, we employ a gradient-reversal layer to invert the gradient emanating from the OE-classifier for all layers in the shared ResNet18 stem. We tuned DR's gradient-reversal hyperparamter λ by grid-search to optimize validation accuracy on Y, while minimizing performance on C.

StarGANv2: We assess StarGANv2 (SG2) [5] as a SOTA style-transfer method in natural images. We train using default parameters over 75k iterations. We sample content and style image pairs as for IST and use OE-labels $\mathbf{Y}$ as domains in SG2's multi-task discriminator. Following training, we use SG2 in the same way as IST, to project input images to random OEs, in the hope to yield an interventional training distribution free of spurious correlations.

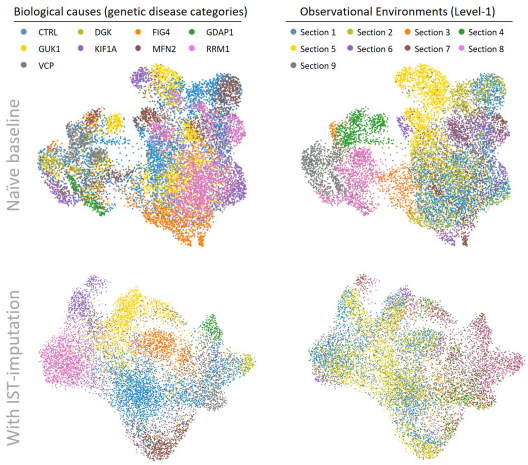

Figure 5. UMAPs of GRID-data training-set features extracted from the penultimate layer of predictors trained with or without IST. Colors show disease-categories Y (left) and OEs C (right)

MixStyle: Finally, we assess MixStyle (MS) [7] as a second recent style-transfer baseline that was specifically developed for domain generalization. We implement MS in our predictor architecture and successfully recapitulate [7]'s results on PACS using our setup. For fairer comparison to IST, we also test MS in a domain adaptation setup, in which we allow MS to train on styles (but not biological causes $\mathbf{Y}$) of images from test OEs.

5.3. Evaluation metrics

OOD-generalization: To test OOD-generalization, we perform section-, plate-, or series-wise cross-validation (*levels of generalization*, see Fig. 3) by testing predictors on OEs that were left out during training.

UMAPs: While qualitative, feature-space visualizations are widely used in the biomedical literature. We report Uniform Manifold Approximation and Projection (UMAPs) [25].

LISI-score: We use the Local Inverse Simpson's Index (LISI) [19] to quantify the local diversity in feature space: in causal representations, we would expect $\mathbb{E}_C \sum_{c \in C}^{C} p(c) = 1$ while $\mathbb{E}_Y \sum_{y \in Y}^{Y} p(y) = 1/|y|$.

kNN-CV: As a second feature-space based metric, we simulate our OOD-generalization experiments by evaluating kNN-classifiers on predicting cause-labels for validation-set images from OEs the corresponding training set images of which are left out of the kNN-reference set.

	GRID						LINCS-SC				
	IID	cLISI/bLISI	kNN BatchCV	Level-1	Level-2	Level-3	IID	kNN BatchCV	Level-1	Level-2	Level-3
Baseline	0.55	0.5417	0.4458	0.1877	0.1381	0.1254	**0.57**	0.1271	NA	0.4194	0.0405
Baseline kNN	0.63	0.5417	0.4458	0.1838	0.1378	0.1317	0.55	0.1271	NA	0.3897	0.0452
Symphony kNN	0.50	0.7340	0.4404	0.1797	0.1474	0.1325	0.35	0.1098	NA	0.2697	0.0461
DR $\alpha = 0.0625$	**0.73**	1.0811	**0.6906**	0.1900	0.1379	0.1259	0.56	0.1381	NA	0.4256	0.0438
MixStyle-DA	0.60	0.6039	0.5519	0.1084	nc	nc	0.41	nc	NA	0.4250	0.0450
StarGANv2	0.20	1.099	0.1539	0.1659	0.1284	0.0977	nc	nc	NA	nc	nc
IST (ours)	0.60	**1.4963**	0.5815	**0.5839**	**0.5350**	**0.3673**	0.53	**0.3304**	NA	**0.7016**	**0.3138**

Table 1. Macro f1-scores on predictor performance on GRID and LINCS-SC. We report kNN-based classification results for Symphony as predictor outputs are not available for post-hoc correction methods. Level-1 cannot be computed for LINCS-SC. nc: not computed

6. Results

We report empirical results for IST and all baselines for GRID and validate our results on LINCS-SC (Table 1). Trained across all OEs, our naive baseline achieves excellent performance on IID hold-out data, suggesting there exist robust phenotypic manifestations of inherent genetic (GRID) and pharmacological (LINCS-SC) causes in the observed single-cell images. However, visual inspection of the resulting feature spaces via UMAP reveals OEs as a prominent superstructure in our models' representations, whereas biological causes form secondary clusters within the local context of their parent OE (Fig. 5). Consistently, LISI scores indicate poor integration over OEs and performance deteriorates in kNN-CV. Critically, when tested on OOD-generalization, our naive baseline shows almost complete collapse across all three levels of generalization on GRID and level-3 for LINCS-SC.

As expected, we find that SYM excels at purging variation over OEs when assessing LISI-scores on the training set. However, for both GRID and LINCS-SC data, we find that this effect does not generalize even to IID validation data and performs poorly on all other metrics. We find that DR-models achieve LISI-scores similar to SYM on GRID, while fully generalizing to IID data, and drastically improving kNN-CV scores. On LINCS-SC however, DR yields only comparatively minor improvements in kNN-CV scores. Remarkably, despite these somewhat promising auxiliary metrics, DR does not significantly improve OOD-generalization across any level in either dataset. Likely because SG2 permutes both style and content (see Fig. 2), predictors trained on the SG2-generated distribution fail even at IID generalization. MixStyle on the other hand, yields excellent IID-performances but - presumably hampered by it's assumptions about what constitutes style-features - yields equally disappointing results in our OOD tests.

By contrast, we find that IST learns to faithfully impute observations as if they had been made in different OEs (Fig. 4B). Qualitative inspection of output images suggests that IST simultaneously preserves phenotypic content of the source images (Fig.4C). As such, IST is able to randomize over the confounder C (Fig.4B) to yield a training distribution $P(y, \hat{x}, z|do(c))$ in which the original correlations between OEs and biological causes are diminished. Consistent with this, we observe major performance gains across all levels of OOD-generalization, as well as other metrics, for both GRID and LINCS-SC data, when predictors are trained on IST-generated data-distributions (Table 1). These results suggest that our IST approach generates effective interventions on confounders and thereby promotes the emergence of causal representations of biological phenomena.

7. Conclusions

Learning visual features that generalize across environments is a critical prerequisite for real-world applications of machine learning systems in biomedicine, yet the field lacks broadly adopted metrics to assess progress towards this goal. We propose OOD-generalization tests structured according to a hierarchy of technical processing steps that generally characterize the data generation process of most high-content imaging studies. We show that seemingly well-performing baselines, including SOTA-methods for batch-effect correction, as assessed by IID hold-out sets and several auxiliary metrics, almost completely collapse on this benchmark, revealing highly confounded representations. The success of IST instead shows that effective interventions to mitigate confounders can be learned, given they are at least partially observed. We point out that even models trained on billions of diverse natural images have only achieved minor gains on ObjectNet [13], suggesting that scale alone is not efficient at breaking contextual biases. Conversely, we suggest our approach bears semblance to thought experiments, by which humans routinely reevaluate familiar concepts in never-observed contexts, thus filling in a sparse matrix of actual observations. We propose IST as a fruitful direction for efficient causal learning.

7.1. Acknowledgements

This research is based on work partially supported by Muscular Dystrophy Association (MDA) Development Grant 628114, and NIH award K99HG011488 to WMP, and by the Broad Institute Schmidt Fellowship program (JCC) and by NSF award 2134695 to JCC.

References

[1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 1

[2] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016. 6

[3] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pages 2023–03, 2023. 1

[4] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: unified generative adversarial networks for multi-domain image-to-image translation. corr abs/1711.09020 (2017). *arXiv preprint arXiv:1711.09020*, 2017. 3

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 3, 4, 7

[6] Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, et al. The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine*, 23(4):405–408, 2017. 6

[7] Zhou et. al. Domain generalization with mixstyle. *ICLR 2021*. 3, 4, 7

[8] Marta M Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023. 1

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 7

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[11] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021. 3

[12] Peter Goldsborough, Nick Pawlowski, Juan C Caicedo, Shantanu Singh, and Anne E Carpenter. Cytogan: generative modeling of cell images. *BioRxiv*, page 227645, 2017. 2

[13] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022. 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[17] Joyce B Kang, Aparna Nathan, Kathryn Weinand, Fan Zhang, Nghia Millard, Laurie Rumker, D Moody, Ilya Korsunsky, and Soumya Raychaudhuri. Efficient and precise single-cell reference atlas mapping with symphony. *Nature communications*, 12(1):1–21, 2021. 3, 7

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[19] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019. 7

[20] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 3

[21] Daniel D Lee, P Pham, Y Largman, and A Ng. Advances in neural information processing systems 22. Technical report, Tech. Rep., Tech. Rep, 2009. 2

[22] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 3, 5

[23] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 5

[24] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021. 1, 2, 3

[25] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[26] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A Cimini, Kyle W Karhohs, Minh Doan, Liya Ding, Susanne M Rafelski, Derek Thirstrup, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, 16(7):e2005970, 2018. 7

[27] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Claire McQuin, Matthew Smith, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022. 7

[28] Asher Mullard. Machine learning brings cell imaging promises into focus. *Nat. Rev. Drug Discovery 2019*. 1

[29] Judea Pearl. *Causality*. Cambridge university press, 2009. 1, 2, 3

[30] Wesley Wei Qian, Cassandra Xia, Subhashini Venugopalan, Arunachalam Narayanaswamy, Michelle Dimon, George W Ashdown, Jake Baum, Jian Peng, and D Michael Ando. Batch equalization with a generative adversarial network. *Bioinformatics*, 36(Supplement_2):i875–i883, 2020. 3

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 3

[32] Lauren Schiff, Bianca Migliori, Ye Chen, Deidre Carter, Caitlyn Bonilla, Jenna Hall, Minjie Fan, Edmund Tam, Sara Ahadi, Brodie Fischbacher, et al. Integrating deep learning and unbiased automated high-content screening to identify complex disease signatures in human fibroblasts. *Nature Communications*, 13(1):1590, 2022. 2

[33] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022. 1, 3

[34] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 1

[35] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020. 2

[36] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11:1–8, 2017. 2

[37] Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo, Beth A Cimini, Kyle Karhohs, David J Logan, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Systems*, 2022. 6

[38] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 2

[39] Karren Yang, Samuel Goldman, Wengong Jin, Alex X Lu, Regina Barzilay, Tommi Jaakkola, and Caroline Uhler. Mol2image: Improved conditional flow models for molecule to image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6688–6698, 2021. 2

[40] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. 3

[41] 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. 5