Dealing with severe class imbalance poses a major challenge for many real-world applications, especially when the accurate classification and generalization of minority classes are of primary interest. In computer vision and NLP, learning from datasets with long-tail behavior is a recurring theme, especially for naturally occurring labels. Existing solutions mostly appeal to sampling or weighting adjustments to alleviate the extreme imbalance, or impose inductive bias to prioritize generalizable associations. Here we take a novel perspective to promote sample efficiency and model generalization based on the invariance principles of causality. Our contribution posits a meta-distributional scenario, where the causal generating mechanism for label-conditional features is invariant across different labels. Such causal assumption enables efficient knowledge transfer from the dominant classes to their under-represented counterparts, even if their feature distributions show apparent disparities. This allows us to leverage a causal data augmentation procedure to enlarge the representation of minority classes. Our development is orthogonal to the existing imbalanced data learning techniques thus can be seamlessly integrated. The proposed approach is validated on an extensive set of synthetic and real-world tasks against state-of-the-art solutions.

1 Introduction

Learning with imbalanced datasets is a common yet still very challenging scenario in many machine learning applications. Typical scenarios include: (i) rare events, where the event prevalence is extremely low while their implications are of high cost, e.g., severe risks that people seek to avert [55]; (ii) emerging objects in a dynamic environment, which call for quick adaptation of an agent to identify new cases with only a handful target examples and plentiful past experience [30]. A typical scenario in natural datasets is that the occurrence of different objects follows a power law distribution. And in many situations, the accurate identification of those rarer instances bears more significant social-economic values, e.g., fraud detection [23], driving safety [26], nature conservation [49], social fairness [10], and public health [62, 91].

Notably, severe class imbalance and lack of minority labels are the two major difficulties in this setting, which render standard learning strategies unsuitable [56]. Without explicit statistical adjustments, the imbalance induces bias towards the majority classes. On the other hand, the lack of minority representations prevents the identification of stable correlations that generalize in predictive settings. In addition, due to technological advancements, more and higher dimensional data are routinely collected for analysis. This also inadvertently exacerbates the issue of minority modeling as there is an excess of predictors relative to the limited occurrence of minority samples.

*Contributed equally.
Various research efforts have been directed to address the above issues, with class re-balancing being the most popular heuristic. The two most prominent directions in the category include statistical resampling and reweighting. Resampling alters the exposure frequency during training, i.e., more for the minorities and less so for the majorities [28]. Alternatively, reweighting directly amends the relative importance of each sample based on their class [21], sampling frequency [7] and associated cost for misclassification [29]. Recent developments also have considered class-sensitive and data-adaptive losses to more flexibly offset the imbalance [10, 59, 64, 79, 50]. While being intuitive and working reasonably well in practice, an important is that these approaches offer no protection against the over-fitting of minority instances (see Figure 1), a fundamental obstacle towards better generalization.

To circumvent this key limitation, inductive biases are often solicited to impose strong constraints that suppress spurious correlations that hinder out-of-sample generalization. A classical strategy is few-shot learning [30, 88, 39], where the majority examples are used to train meta-predictors and transferable features, leaving only a few parameters to tune for the minority data. In anomaly detection, methods such as one-class classification instead regard minority classes as outliers that do not always associate with stable, recognizable patterns [66, 74].

Despite their relatively strong assumptions, these methods capitalize on their superior ability in generalizing in the low sample regime in empirical settings.

More recently, establishing causal invariance has emerged as a new, powerful learning principle for better generalization ability even under apparent distribution shifts [72]. In contrast to standard empirical risk minimization (ERM) schemes, where the generalization to similar data distributions is considered, causally-inspired learning instead embraces robustness against potential perturbations [5]. This is achieved via only attending to causally relevant features and associations postulated to be invariant under different settings [69]. Specifically, contributions from spurious, unstable features are effectively blocked or attenuated. Interestingly, compromise of performance can be expected in those models [73], as a direct consequence of discarding useful (but non-causal) correlations in exchange for better causal generalization. Recently, [47] proposed non-parametric causal disentanglement of data representation to identify invariant relations.

This work explores the advancement of imbalanced data learning via adopting causal perspectives, with the insight that the key algorithm can be reformulated as an energy-based contrastive learning procedure to drastically improve efficiency and flexibility. In recognition of the limitations discussed above, we present energy-based causal representation transfer (ECRT): a novel imbalanced learning scheme that brings together ideas from causality, contrastive learning, energy modeling, data-augmentation and weakly-supervised learning, to address the identified weakness of existing solutions. Our key contributions are: (i) a causal representation encoder informed by an invariant generative mechanism based on generalized contrastive learning; (ii) integrated data-augmentation and source representation regularization techniques exploiting feature independence to enrich mi-
nority representations that better balance the trade-off between utility and invariance; (iii) a key novelty is the derivation of an energy-based contrastive learning algorithm that greatly enhance model parallelism for large-label settings and extends generalized contrastive learning; and (iv) insightful discussions on the justifications for the use of the proposed approach. Our claims are supported by strong experimental evidence.

2 Preliminaries

Notation and problem definition. We use \( x \in \mathbb{R}^p \) to denote the input data and \( z \in \mathbb{R}^d \) for the predictive features extracted from \( x \). Let \( y \in \{1, \ldots, M\} \) be the class label. The number of training samples and those with label \( m \) are denoted as \( n \) and \( n_m \), respectively. We use \( \mathbb{E}_x[\cdot] \) to denote the expectation (average) of an empirical distribution, \( a_i \) to denote a vector associated with the \( i \)-th sample, and \( a_{b_i} \) to denote the \( b \)-th entry of vector \( a \). For simplicity, we assume class \( M \) is the minority class, i.e., \( n_m \gg n_M \), for \( m \in \{1, \ldots, M - 1\} \). Throughout, we refer to \( X, Z \) and \( S \) as data, feature, and source spaces, respectively, as shown in Figure 2, and defined below. Our goal is to accurately predict the label of minority instances with very limited training examples of it. Generalization to multiple minority categories is straightforward.

Generalized contrastive learning and ICA. The proposed approach is based on a generalized form of independent component analysis (ICA), which addresses the inverse problem of signal disentanglement [45]. Specifically, ICA decorrelates features \( Z \) of the observed signal \( X \) into a source signal representation \( S = f_{\psi}(Z) \), where \( f_{\psi}(z) \) is a smooth and invertible mapping known as the de-mixing function, and while assuming that the components of \( S \) are statistically independent, i.e., with density \( q(\mathbf{s}) = \prod_j q_j([s]_j) \). Notationally, we call \([s]_j\) the \( j \)-th independent component (IC) of \( Z \). While nonlinear ICA (NICA) is generally infeasible [18][46], [47] has recently proposed a setting in which the identification of NICA can be achieved, by requiring an additional auxiliary label \( y \). Specifically, NICA assumes that source signals are conditionally independent given \( y \), i.e., \( q(\mathbf{s}|y) = \prod_j q_j([s]_j|y) \), then \( f_{\psi}(z) \) can be identified using generalized contrastive learning (GCL), whose implementation is detailed below.

GCL solves a generalized regression problem in which a critic function predicts whether label \( y \) and representation \( z \) are correctly paired, i.e., congruent. We call \((y, z)\) a congruent pair and \((y, z)\) an incongruent pair if \( i \neq j \). Specifically, the critic is defined as \( r_{\nu}(y, z) = \sum_{a=1}^d r_{\nu}^a(y, [s]_a) \), where \( r_{\nu}^a(\cdot, \cdot) \) is a neural network with parameters \( \nu \), whose inputs are the label \( y \) and the \( a \)-th coordinate of \( s = f_{\psi}(z) \), denoted as \([s]_a\). Then, GCL optimizes the following objective:

\[
\arg\min_{f_{\psi}, \nu_{\nu}} \mathbb{E}_x[h(-r_{\nu}(y, z_i)) + \mathbb{E}_{j \neq i}[h(r_{\nu}(y, z_i))]],
\]

where \( h(r) = \log(1 + \exp(r)) \) is the softplus function. (1) seeks to optimize \( f_{\psi}(\cdot) \) and \( r_{\nu}(\cdot, \cdot) \) by maximizing the discriminative power to tell apart congruent and incongruent pairs.

In fact, an interesting result showed by [47] revealed that maximizing the ability of the critic function \( r_{\nu}(\cdot, \cdot) \) for correctly identifying matching pairs \((z, y)\) leads to the identification (up to univariate transformation) of \( f_{\psi}(\cdot) \), such that the components of \( s \) are conditionally independent given \( y \). See the Supplementary Material (SM) for a formal exposition.

3 Energy-based Causal Representation Transfer

In this section, we describe the construction of energy-based causal representation transfer (ECRT): a causally informed data transformation and augmentation procedure to improve learning with imbalanced datasets. Our model assumes a shared causal data-generation procedure, which can be accurately identified by learning with the majority classes under assumed class-conditional representation independence. We obtain decorrelated representations that facilitate data augmentation and efficient learning with minority classes.
The proposed model consists of the following components: (i) a feature encoder module \( z = e_\theta(x) \); (ii) two classification modules, \( h_\psi(z) \) and \( h_\phi(s) \), for predicting label \( y \) from features \( z \) and sources \( s \), respectively; (iii) a nonlinear ICA module for the de-mixing function \( s = f_\psi(z) \); (iv) a critic function \( r_\nu(y, z) \) for GCL; and (v) a data augmentation module. Further, \( (\theta, \phi', \psi, \nu) \) denote the parameters of all the neural-network-based modules. Algorithm 1 outlines a general workflow, and below, we elaborate on our assumptions and detail the implementation of ECRT.

### 3.1 Model assumptions

To enable knowledge transfer across classes, we make the following assumptions:

**Assumption 3.1.** Let \( z \) be a sufficient statistic (features) of \( x \) for predicting label \( y \), all class conditional feature distributions \( p(z|y) \) share a common ICA de-mixing function \( f_\psi(z) \).

This implies there exists a smooth invertible function \( f_\psi : Z \rightarrow S \), and a set of IC distributions \( \{q(s|y = m)\}_m \) that are linked to the conditional feature distributions \( p(z|y = m) \) via \( S^m = f_\psi(Z^m) \), where \( S^m \sim q(s|y = m) \) and \( Z^m \sim p(z|y = m) \). The subscript \( m \) in \( S^m \) and \( Z^m \) indicates \( y = m \), in a slight abuse of notation.

Importantly, \( f_\psi^{-1}(s) : S \rightarrow Z \) is the **invariant causal mechanism** underlying the features of observed data. This specification, which is consistent with Assumption 3.1, enables the identification of the shared generating process, namely, the de-mixing function \( f_\psi(z) \) that connects the likely dissimilar conditionals \( \{q(z|y = m)\}_m \) via the source conditionals \( \{q(s|y = m)\}_m \), as demonstrated in the context of NICA by [47].

A shortcoming of Assumption 3.1 is that the hypothesis supporting it is somewhat strong, untenable and may not hold in practice. However, we argue that the structural constraints imposed on the model by the assumption via the source conditionals \( q(s|y = m) \) and the invertibility of \( f_\psi(z) \), restrict the search space of the otherwise over-flexible model space powered by neural networks. Further, the causal mechanism implied by \( f_\psi^{-1}(s) \) enables an effective knowledge transfer mechanism across classes via \( q(s|y = m) \).

### 3.2 Energy-based Causal Representation Transfer

**Encoder pre-training.** To implement ECRT, we first find a good (reduced) feature representation of \( z \) highly predictive of label \( y \). This can be achieved via supervised representation learning (see Figure 3), which optimizes an encoder and predictor pair \( (e_\theta(x), h_\psi(z)) \) to minimize the label prediction risk \( \mathcal{L}(\phi') \triangleq \mathbb{E}[\ell(h_\psi(e_\theta(x)), y)] \), where \( \ell(\cdot, \cdot) \) is a suitable loss function, e.g., cross-entropy, hinge loss, etc. To avoid capturing spurious (non-generalizable) features that overfit the minority class, we advocate training only with majority samples at this stage; assuming that \( M > 2 \). Alternatively, one could also consider unsupervised feature extraction schemes, such as auto-encoders. Further, we also recommend using statistical adjustments such as importance weighting, to reduce the impact of data imbalance.

---

**Algorithm 1** Energy-based Causal Representation Transfer.

1. Pre-train encoder and predictor:
   \[
   e_\theta, h_\psi \leftarrow \text{arg} \min \{\mathbb{E}[\ell(h(e_\theta(x)), y)] \}
   \]

2. NICA estimation with fixed \( z = e_\theta(x) \):
   \[
   f_\psi, r_\nu \leftarrow \text{GCL}((z, y)) \quad \text{% Equation (1)}
   \]

3. Source space augmentation with fixed \( s = f_\psi(z) \):
   \[
   (\hat{s}, \hat{y}) \leftarrow \text{Aug}((s|y = M)) \quad \text{% Equation (2)}
   \]

4. Minority predictor modeling with augmented source:
   \[
   h_\phi \leftarrow \text{arg} \min \{\mathbb{E}[\ell(h(s, y = M))] + \lambda \mathbb{E}[\ell(h(\hat{s}, y = M))] \}
   \]
De-mixing representation with GCL. After obtaining a good feature representation \( Z = e_\theta(X) \), we proceed to learn the de-mixing function \( f_\psi(z) \), such that the coordinates of the source representation \( S = f_\psi(Z) \) are (approximately) independent given the label \( y \). This can be done by optimizing the GCL objective in (1) with respect to the feature representation and label pairings \((y, z)\), adopting the \textit{masked auto-regressive flow} (MAF) \cite{papamakarios2017masked} to model the smooth, invertible transformation \( f_\psi(z) \), which allows efficient parallelization of the autoregressive architecture via causal masking. The procedure is outlined in Figure 3.

Augmenting the minority. Inspired by \cite{zhao2017energy}, we artificially augment the minority feature representations \( z \) via random permutations in the source space \( S \), as shown in Figure 4. Provided the assumed conditional independence of the sources, the features \( Z^m \) corresponding to label \( y = m \) are generated by \( Z^m = f_\psi^{-1}(S^m) \), where each dimension in the source representation \( [S^m]_j \sim q([s]_j | y = m) \) are independently and implicitly sampled as described below. Specifically, using the (estimated) de-mixing function \( f_\psi(z) \), we can obtain an approximate empirical source distribution for each \( y = m \), i.e., \( S^m \triangleq \{ s_i = f_\psi(e_\theta(x_i)) | y_i = m \} = \{ s_i \}_{i=1}^{n_m} \), where \( S^m \) is a collection of \( n_m \) samples of \( q(s|y = m) \). Then, we can draw new artificial samples \( \tilde{s}_m \sim q(s|y = m) \) by randomly permuting the coordinates within elements \( S^m \) independently via

\[
\tilde{s}_m = ([s_m]_1, [s_m]_2, \ldots, [s_m]_d),
\]

where \( \sigma = (\sigma_1, \cdots, \sigma_d) \) is a random permutation of \((1, \ldots, n_m)\). Note we have used \( \tilde{s}_m \) to emphasize that the source point is artificially created via permutation \( \sigma \). We call this procedure \textit{nonparametric augmentation} because it does not make distributional assumptions for \( q(s|y = m) \). However, below we will discuss its limitations and consider an alternative where a parametric form is assumed. While it is tempting to refine the predictor \( h_\phi(z) \) with artificial features augmented via \( \tilde{s}_m = f_\psi^{-1}(\tilde{s}_m) \), in Section 3.3 we will argue that it stands to benefit more from training a new predictor directly based on source representations, i.e., \( h_\phi(s) \), without the need for inverting \( f_\psi(z) \).

Model refinement. Now we can leverage the augmented data to refine the prediction model. For minority class \( y = M \), we optimize the following objective

\[
\mathcal{L}_{\text{Aug}}(\phi^\prime) = \mathcal{L}(\phi^\prime) + \lambda \mathbb{E}_Z[\ell(h_\phi(\tilde{Z}^M), M)] - \mathbb{E}_Z[\ell(h_\phi(z), M)],
\]

where \( \mathcal{L}(\phi^\prime) \) is the loss used for pre-training. Conceptually, (3) replaces a portion of the minority samples with augmentations. The trade-off parameter \( \lambda \in [0, 1] \) encodes the relative confidence for trusting the artificially generated representations \( \tilde{Z}^M \) obtained from \( S^M \) for the minority label \( y = M \). Further, at this stage we found it’s beneficial to fix the encoder module to prevent the de-mixing function to accommodate the changes in the encoder which in practice may cause instability during training.

Challenges with naïve implementation. We identify three major issues with naïve implemented ECRT, to be addressed in the section below: (i) \textit{Representation conflict}: since the GCL solution is not unique, we do observe naïve GCL training drifts among viable source representations whose performance differ considerably, causing stability concerns; (ii) \textit{Costly augmentation}: MAF inversions dominate the computation load during training, which becomes prohibitive in high dimensions; and (iii) \textit{Gridding artifact}: a small minority sample size leads to pronounced augmentation bias when sampling nonparametrically via (1), manifested as a rectangular-shaped grid (see Figure 5).

3.3 Improving causal representation transfer

Energy-based GCL. Our key insight to improve GCL comes from the fact that Equation (1) is essentially learning the density ratio between the joint and product of marginals, i.e., \( \frac{p(x,y)}{p(x)p(y)} \). This immediately reminds us the recent literature on contrastive \textit{mutual information} (MI) estimators, such as InfoNCE \cite{chen2020exploring}. In such works, a variational lower bound of MI is derived, and the algorithm optimizes a critic function using the positive samples from the joint distribution, and the negative samples from the product of the marginals. At their optimal value, these critics recover the density ratio or a transformation of it. Our development is based on the recent work of \cite{hjelm2018learning}, using an energy-perspective to improve contrastive learning. Specifically, we will be using a variant of the celebrated \textit{Donsker-Varadhan} (DV) estimator \cite{donsker1983asymptotic}, and applied Fenchel duality trick to compute a solution \cite{vempalawat2016fenchel, nowozin2016fused, shazly2016energy}. Specifically, the \textit{Fenchel-Donsker-Varadhan} (FDV) estimator takes the following form:

\[
I_{\text{FDV}} \triangleq \hat{I}_{\text{DV}}([x_i, y_i]) + \sum_i \exp\left[ (g_\theta(x_i, y_j) - g_\theta(x_i, y_i)) / \tau \right] / \sum_j \exp\left[ (g_\theta(x_i, y_j) - g_\theta(x_i, y_i)) / \tau \right] + 1,
\]
where \( g_0(x_i, y_i) \) is our critic of interest and 
\[
\tilde{I}_{DV}^K(\{x_k, y_k\}) = g(x_1, y_1) - \log(\sum_{k=1}^K \exp(g(x_1, y_k')) / K)
\]

is the Donsker-Varadhan (DV) estimator \([27]\) for the MI. Using the same parameterization used in GCL recovers the same causal identification property (see Appendix for details). Compared to the original GCL formulation, we are now using multiple negative samples instead of one, which greatly boosts learning efficiency \([37]\). And this can be efficiently implemented with the bilinear critic trick \([14, 13, 36]\) so that all in-batch data can be used as negatives. In our context, it greatly boosts training efficiency when dealing with a large number of different classes. See Algorithm S1 in Appendix.

**Regularizing the data likelihood.** Recall GCL solutions can only be identified up to an invertible transformation of each dimension \([47]\), and the predictive performance of different valid GCL solutions can vary significantly. Empirically, we observe that a naïve implementation of GCL often leads to source representations that are densely packed (see Figure S1 in the SM). This is undesirable, when decoding back to the feature space and making useful predictions, the neural network predictor will need to be expansive, i.e., requiring a large Lipschitz constant, thereby sacrificing optimization stability and model generalization according to existing learning theory \([19, 86]\).

To encourage source representations that are less condensed, we consider a simple, intuitive strategy consisting of regularizing the source representation with the log-likelihood in feature space. This likelihood can be easily obtained with MAF using \( \ell_{FLOW}(f_\psi) \) defined in the SM. Following common practice, we set source prior \( p(s) \) to the standard Gaussian, and optimize the following likelihood-regularized GCL objective:

\[
\hat{L}_{GCL}(f_\psi, r_\nu) = L_{GCL}(f_\psi, r_\nu) + \rho L_{FLOW}(f_\psi),
\]

where \( \rho > 0 \) is the regularization strength. Naturally, this regularization will encourage the global source representation identified by GCL to be more consistent with a Gaussian-shaped distribution.

An alternative interpretation for the likelihood-regularized objective in (5) is that it can be understood as a relaxation to the conditional independence (Assumption 3.1). To see this, recall the likelihood objective obtained from the invertible neural networks (INN) alone attempts to map the source representations to be unconditionally independent, as opposed to the conditional independence assumed by NICA. The regularized formulation [3] provides a safe “fall-back” mode in case Assumption 3.1 is violated. This also motivates us to consider an important variant: making multiple class-dependent source priors, i.e., \( p^m(s) \) for each class label \( m \in \{1, \cdots, M\} \) in [3], whose parameters (i.e., mean and variance) are jointly learned with other model parameters. Compared to the fully non-parametric objective [1], this strategy further encourages the source representations to be independent given the class labels, and it enables parametric data augmentation, i.e., sampling from the parametric label priors instead of permuting the indices. We refer to this variant as ECRT-MP, where MP stands for multiple priors. And similarly, ECRT-IP refers to the case when a single prior is used. We have found that ECRT-MP performs better in most cases, and consequently, ECRT means ECRT-MP by default.

**Modeling in the source space.** Rather than modeling the predictor \( h_\phi(z) \) in the feature space \( Z \), we advocate instead for building the predictor directly in the source space \( S \), i.e., modeling with \( h_\phi(s) \). This practice enjoys several benefits: (i) **Easy & robust augmentation:** many designs of high-dimensional flows are asymmetric computationally, and inverting a MAF is not only \( d \) times more costly than a forward pass, it is also numerically unstable at the boundary. Direct modeling in the source space circumvents the difficulties associated with MAF inversions during data augmentation; (ii) **Feature whitening:** the source representation identified by GCL is component-wise independent, and variance (con-ditional) are jointly learned with other model parameters. Compared to the fully non-parametric objective (1), this strategy further encourages the source representations to be independent given the class labels, and it enables parametric data augmentation, i.e., sampling from the parametric label priors instead of permuting the indices. We refer to this variant as ECRT-MP, where MP stands for multiple priors. And similarly, ECRT-IP refers to the case when a single prior is used. We have found that ECRT-MP performs better in most cases, and consequently, ECRT means ECRT-MP by default.

**Parametric augmentation.** When the number of minority observations is scarce, the above non-parametric indices-shuffling augmentation suffers from the gridding artifact (Figure 5). This artifact amplifies the augmentation bias in the low-sample regime. To overcome this limitation, we empirically observe that the estimated class-conditional source distributions are usually Gaussian-like after the
likelihood regularization (especially so when label conditional priors are used). In these situations, a parametric augmentation that draws synthetic source samples from a Gaussian distribution matched to the empirical mean and variance of minority source representations is more efficient.

### 3.4 Insights and remarks

To better appreciate the gains and limitations expected from ECRT, we compile a few complementary arguments below, through the lens of very different perspectives.

**Why causal augmentation works.** It is helpful to understand the gains from ECRT’s causal augmentation beyond the heuristic that permuting the ICs provides more training samples for the minority class. [82] considered a similar causal augmentation procedure for few-shot learning, and provided two major theoretical arguments: (i) the risk estimator based on the augmented examples is the uniformly minimum variance unbiased estimator given the accurate estimation of \( f_\psi \) (see Theorem 1, [82]); and (ii) with high probability, the generalization gap can be bounded by the approximation error of \( f_\psi \) (see Theorem 2, [82]). In the SM, we give arguments that our causal augments give the ‘best’ label-conditional distribution estimate.

**Speedup from shared embedding.** While for typical supervised learning tasks the generalization bound scale as \( O(n^{-\frac{1}{2}}) \), a superior rate of \( O(n^{-\eta}) \) where \( \eta \in \left[\frac{1}{2}, 1\right] \) is possible, if there exists abundant data for an alternative, yet related task that shares the same feature embedding (see Theorem 3, [71]). Note \( n \) refers to the size of labeled data directly related to the strong task of interest, in our case, prediction of minority labels. Our ECRT employs GCL to identify one such common embedding, i.e., the source space, using the majority examples, and consequently, improves predictions on the minority class.

**Representation whitening.** Our ECRT causally disentangles representation [77, 83] via de-correlating the representations conditionally. Extensive empirical evidence has pointed to the fact that such representation de-correlation, more commonly known as data whitening [51], is expected to considerably improve learning efficiency [17]. This benefit has been attributed to the better conditioning of the Fisher information matrix for gradient-based optimization [24], thus rectifying second-order curvatures to accelerate convergence. Our source space modeling explicitly separates the task of representation disentanglement, and in turn, helps the prediction network to focus on its primary goal.

**Potential limitations.** The setting considered by ECRT is restrictive in that it precludes the learning of useful, yet non-transferable features predictive of the minority labels. For instance, there might be a feature unique to the minority class. However, since the de-mixer \( f_\psi(z) \) is only trained on the majority domains absent of this feature, it can not be accounted for by the ECRT model. This is a key limitation of causally inspired models, in that they are often too conservative for only retaining the invariant features, promoting cross-domain generalization at the cost of within-domain performance degradation [73].

### 4 Related Work

**Causal invariance and representation learning.** A major school of considerations for building robust machine learning models is to stipulate invariant causal relations across environments [75, 61, 12], such that one hopes to safely extrapolate beyond the training scenario [8]. We broadly categorize such efforts into two streams, namely predictive causal models and generative causal models. Prominent examples from the first category include ICP [69, 42] and IRM [5], which highlight the identification of invariant representation and causal relations via penalizing environmental heterogeneity. Our solution pertains to the second category, where the data distribution shifts across environments can be tethered by an invariant generation procedure [47, 52]. This work exploits the recovered causally invariant source representation to improve learning efficiency and mitigate the sample scarcity of minority labels in imbalanced sets.

**Learning with imbalanced data.** Resolving data imbalance is a heavily investigated topic [60]. Standard sampling and weight adjustments suffer from caveats such as introducing bias and information loss. Due to these concerns, recent literature has actively explored adaptive strategies, such as redundancy-adjusted balancing weights [21], and the theoretically grounded class-size adapted margins [10]. Similar to boosting [33], adaptive weights are designed to prioritize the learning of less well-classified examples [59] while excluding apparent outliers [58]. Much related to our setting are the meta-learning scenarios [87, 52, 89], where the model tries to generalize & repurpose the
knowledge learned from majority classes to minority predictions; and also the augmentation-based schemes to restore class balance [63, 67].

Data augmentation. Due to its exceptional effectiveness, augmentation schemes are widely adopted in practical applications [76]. These augmentation strategies are built on known invariant transformations, e.g., rotation, scaling, noise corruption, etc. [78], simple interpolation heuristics [11, 40], and more recently towards fully auto-mated procedures [20]. Notably, recent trend in data augmentation highlights robust learning against adversarially crafted inputs [44] and generative augmentation procedures that compose realistic artificial samples [3].

Domain adaptation and causal mechanism transfer. Closest to our contribution is causal mechanism transfer (CMT) [82], which focused on addressing few-shot learning for continuous regression. We note a few key differences to our work: (i) CMT focused on domain adaptation and does not address classification; and (ii) it bundles \((x, y)\) for NICA which necessitates the flow inversion for sample augmentation. Grounded in the setting of imbalanced data learning, our ECRT extends applications and advocates source space modeling to simplify and improve causal augmentation. It also features likelihood regularization to enhance representation regularity. We also offer new insights to justify the use of NICA-based augmentation, complementing the analysis from CMT by [82].

Energy-based modeling for representation learning. There is growing recognition that the energy perspective is integral to representation learning [57, 70, 81, 36]. In the lens of energy based modeling, the distribution of data is characterized as an (unnormalized) energy function [4]. Optimization of the energy function is often considered challenging [80] and contrastive techniques have been proven effective [43, 38]. Our work is a generalization of [42] that disentangles representations from an energy perspective. Interestingly comparison can be made to [53], whose training objective bears resemblance to our ECRT. But the specific designs used by ECRT in the network architecture and scoring function allows ECRT to provably disentangle the feature representations and consequently capture causality for exploitation.

5 Experiments

To validate the utility of our model, we consider a wide range of (semi)-synthetic and real-world tasks experimentally. All experiments are implemented with PyTorch, and our code is available from https://github.com/ZidiXiu/ECRT. More details of our setup & additional analyses are deferred to the SM Sections C-E.

5.1 Experimental setup

Baselines. The following competing baselines are considered to benchmark the proposed solution: (i) Empirical risk minimization (ERM), a naive baseline with no adjustment; (ii) Importance-weighting (IW) [9], a class-weight balanced training loss; (iii) Generative adversarial augmentation (GAN) [3], synthetic augmentations from adversarially-trained sampler; (iv) Virtual adversarial training (VAT) [65], robustness regularization with virtual perturbations; (v) FOCAL loss [59], cost-sensitive adaptive weighting; (vi) Label-distribution-aware margin loss (LDAM) [10], margin-optimal class weights. All baselines are tuned for best performance for Top-1 accuracy on validation.

Evaluation metrics & setup. We consider the following metrics to quantitatively assess performance: (i) negative log-likelihood (NLL); (ii) F1 score; (iii) Top-k accuracy \((k = 1, 5)\). Following the classical evaluation setup for imbalanced data learning, we learn on an imbalanced training set and report performance on a balanced validation set.
Toy model. We sample 2D standard Gaussians with different means and variances as source representation $s$ for each label class, which is then distorted by random affine and Hénon transformations $z_i = 1 - 1.4 s_i^2 + s_2, z_2 = 0.3 s_1, i \in [1, 2]$, to induce real-world-like complex association structures. See the SM Sec. D for a detailed description and visualizations.

Real-world datasets. We consider the following semi-synthetic and real datasets: (i) Imbalanced MNIST and CIFAR100: standard image classification tasks with artificially created step-imbalanced following [10]; (ii) Imbalanced TinyImageNet [2]: a scaled-down version of the classic natural image dataset ImageNet, comprised of 200 classes, 500 samples per class and 10k validation images, with different simulated imbalances applied; (iii) iNaturalist 2019 [85]: a challenging task for image classification in the wild comprised of 26k training and 3k validation images labeled with 1k classes, with a skewed distribution in label frequency; (iv) arXiv abstracts, imbalanced multi-label prediction of paper categories with 160k samples and 152 classes.

Preprocessing, model architecture, and tuning. We use pre-trained models to extract vectorized representations for big complex datasets: ResNet [41] for image models of TinyImageNet [2], Inception V3 for iNaturalist [4] and BERT [25] for the arXiv [11] language model. These pre-trained representations are used as raw feature inputs, subsequently fed into a fully-connected multi-layer perceptrons (MLP) for source space encoding. For small image models (i.e., MNIST, CIFAR100) we directly train a CNN from scratch for feature extraction. We used a random 8/2 split for training and validation, and applied Adam optimizers for training. We rely on the best out-of-sample cross-entropy and GCL loss for hyperparameter tuning.

ECRT efficiency. We first examine the dynamics of efficiency gains from ECRT using a toy model, with the results summarized in Figure 9. Consistent with the shared embedding perspective, increasing the majority sample size improves the minority accuracy. Causal augmentation consistently improves performance, and it is most effective in the low-sample regime, offering more than $4 \times$ boosts in effective minority sample size. In Figure 10 we compare modeling with feature and source representations, respectively. Source space modeling speeds up learning, in addition to the modest accuracy gains. Augmented training is slightly slower, but eventually convergences to a better solution.

Table 1: Comparison of performance on real-world datasets (↑ higher is better, ↓ lower is better).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR100</th>
<th>iNaturalist</th>
<th>TinyImageNet</th>
<th>ArXiv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1↑</td>
<td>Top-5↑</td>
<td>NLL↓</td>
<td>Top-1↑</td>
</tr>
<tr>
<td>ERM</td>
<td>49.29</td>
<td>78.32</td>
<td>2.95</td>
<td>66.73</td>
</tr>
<tr>
<td>IW</td>
<td>43.97</td>
<td>68.96</td>
<td>3.89</td>
<td>67.63</td>
</tr>
<tr>
<td>GAN</td>
<td>47.64</td>
<td>78.53</td>
<td>2.69</td>
<td>67.40</td>
</tr>
<tr>
<td>VAT</td>
<td>46.47</td>
<td>74.23</td>
<td>3.19</td>
<td>67.06</td>
</tr>
<tr>
<td>FOCAL</td>
<td>43.52</td>
<td>74.39</td>
<td>2.89</td>
<td>66.63</td>
</tr>
<tr>
<td>LDAM</td>
<td>50.46</td>
<td>74.39</td>
<td>2.18</td>
<td>67.39</td>
</tr>
<tr>
<td>ECRT-1P</td>
<td>52.31</td>
<td>81.14</td>
<td>1.98</td>
<td>68.38</td>
</tr>
<tr>
<td>ECRT-MP</td>
<td>53.00</td>
<td>81.99</td>
<td>2.31</td>
<td>69.01</td>
</tr>
</tbody>
</table>

https://download.pytorch.org/models/resnet18-5c106cde.pth
https://download.pytorch.org/models/inception_v3_google-1a9a5a14.pth
https://github.com/allenai/scibert
Interestingly, the major performance gain originates from the GCL-based source space modeling we proposed, and to a lesser extent, from the augmentation perspective presented by [82].

**Ablations on augmentation.** We examined the contributions from augmentation as we vary the augmentation strength $\lambda$ (Figure 11). A relatively small $\lambda$ already yields improvement, but when the minority size is small, stronger augmentation degrades performance. This implies the augmentation gains are more likely to be originated from the exposure to the diversity of synthetic augmenting, and the accuracy of augmented distribution is limited by the imperfections of empirical estimation (e.g., limited sample size). In Figure 7, we compare the MMD distance [35] between the parametric and nonparametric augmentation outputs and the ground-truth, confirming the improved efficiency from the ECRT parametric augmentation in the low-sample regime.

**Ablation on different model variants.** We further compare the model performance with feature encoder trained with both majority and minority. (Table 1 used feature encoder trained only majority classes.) We have Cifar100 TOP-1=51.79, TOP-5=81.02, NLL=2.03. This is slightly worse than the majority-only result but still outperformed other competing solutions. We caution, whether including minority examples in the pre-training of feature encoder differs case-by-case. They may not affect performance at all (majority dominant), improve performance (predictive features consistent with those used by majority) and completely devastating (containing spurious features, overfit).

**Extreme classification.** A related challenge of interest is extreme classification [15], where the label size is huge but there is only a handful of samples pertaining to each label class. This means there is no clear majority class. In Figure 6, we show the proposed ECRT also fares very well in this scenario, while standard ERM struggles (slightly better than random guess). This result evidences that ECRT also efficiently extracts generalizable information from an abundance of different class labels.

**Imbalanced data learning.** In Table 1, we compare the performance of ECRT to other competing solutions, with each baseline carefully tuned to ensure fairness. Minor discrepancies compared to results reported by prior literature are explained in the SM Sec. E, and in general our implementation performs slightly better. Overall, ECRT-based solutions consistently lead the performance chart, with the label-prior variant solidly outperforming vanilla ECRT and other competitors in most categories. While showing varying degrees of success, FOCAL and LDAM failed to establish dominance compared to ERM. GAN and VAT also verified the effectiveness of augmentation and adversarial perturbations in the imbalanced setting. We also compare the overall F1-score between different methods on CIFAR100 dataset: ERM: 0.439, SMOTE: 0.444, FOCAL: 0.391, LDAM: 0.408, and ECRT: 0.482. FOCAL and LDAM gave very poor F1 scores in this case, even worse than the ERM baseline. SMOTE improved ERM, but the largest gain is obtained by ECRT. Figure 1 compares performance of different learning schemes condition on different label sizes using the F1 score on the iNaturalist data. In Figure 8, we show the ECRT learned representations, with and without augmentation, using tSNE embeddings.

### 6 Conclusion

This paper developed a novel learning scheme for imbalanced data learning. Leveraging a causal perspective, our solution cuts through the apparent data heterogeneity and identifies a shared, invariant, disentangled source representation, using only majority samples. We demonstrate this source representation modeling enables more efficient learning and allows principled augmentation. Importantly, we bridge the research between contrastive representation learning and energy modeling to causal machine learning, which constructs promising directions for future research.

### Acknowledgements

This research was supported in part by NIH/NIDDK R01-DK123062, NIH/NIBIB R01-EB025020, NIH/NINDS 1R61NS120246, DARPA, DOE, ONR and NSF. J Chen was partially supported by Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01) and National Key R&D Program of China (No.2018AAA0100303). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 [84]. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) PSC Bridges-2 and SDSC Expanse at the service-provider through allocation TG-ELE200002 and TG-CIS210044.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Please check the Supplemental Material.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer to Sec.5 and the SM.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] All datasets are public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The datasets don’t have any personally identifiable information.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]