Supplementary Materials for

# Learning Disentangled Representations for Environment Inference in Out-of-distribution Generalization

**Algorithm 1** VAE-based environment inference invariant learning (invariant learning by IRMv1 penalty)

**Input**: Training dataset $\mathcal{D}_{tr} = \{x_i, y_i\}^N$, training epoch of reference model $N_{ref}$, environment inference step $N_{EI}$, invariant learning step $N_{IL}$

**Parameter**: VAE encoder $\phi$ and decoder $\theta$, classifier $\varphi$, invariant learning predictor $\Phi$

**for** $t \in 1, ..., N_{ref}$ **do**
  Update $\phi, \theta, \varphi$ via Eq. 4
**end for**
Randomly init $q \in [0, 1]^N$
**for** $t \in 1, ..., N_{EI}$ **do**
  loss $= -\left\|\nabla_{\overline{w}} \widetilde{R}^e(\overline{w} \cdot \varphi \cdot \phi, q)\right\|^2$
  Update $q$ via $\nabla_q loss$
**end for**
$\hat{q} = Bernoulli(q)$
$\mathcal{E}_0 \leftarrow \{x_i, y_i | \hat{q}_i = 1\}, \mathcal{E}_1 \leftarrow \{x_i, y_i | \hat{q}_i = 0\}$
**for** $t \in 1, ..., N_{IL}$ **do**
  Update $\Phi$ via Eq. 1 by given $e \in \{\mathcal{E}_0, \mathcal{E}_1\}$
**end for**

## A Psuedocode

Algorithm 1 provides pseudocode for our VAE-based environment inference procedure used in our experiments.

## B Implementation Details

### B.1 Network Implementation

We implement the encoder network on Colored MNIST[I] and Colored MNIST[II] using MLP with three-layer hidden layers. We also use similar MLP architecture in the real-world CelebA dataset, because the environment inference step needs a full-batch gradient descent which takes expensive memory

| | SFS |
|---|---|
| $\beta = 0.1$ | 0.74 |
| $\beta = 0.5$ | 0.81 |
| $\beta = 1$ | 0.77 |
| $\beta = 5$ | 0.72 |
| $\beta = 10$ | 0.73 |
| $\beta = 50$ | 0.64 |
| $\beta = 100$ | 0.47 |

**Table S1:** Spurious feature score of the spurious representation with different values of $\beta$ for the penalty term with KL divergence in VAE.

usage. Moreover, CelebA offers the 512 dimensions features of samples which allow us to use the model with a small size. All baseline methods have the same encoder for a fair comparison.

### B.2 Hyperparameters

We reuse the IRM penalty, L2 regularization weight, and environment inference steps used in IRM and EIIL after model selection by previous authors[1]. For Colored MNIST[II], we increase the environment inference steps to 30000 for the convergence of soft assignment. The dimension of latent variable $dim(z)$ in our VAE-based method are 64, 64, and 128 in the Colored MNIST[I], Colored MNIST[II], and CelebA. Our method has $\lambda$ used for adjusting the VAE term and classifier term which is set to 1.0 for all datasets.

### B.3 Weighted KL Divergence Term

We can notice that VAE has a hyperparameter $\beta$, which is the coefficient on the Kullback-Liebler (KL) divergence term in Eq. 3. The higher $\beta$ encourages a VAE to learn unsupervised disentangled representation. Table S1 shows the impact on spurious features with different values of $\beta$. In our experiments, the higher $\beta$ does not benefit the performance of our model and we set it to 0.5.

### B.4 Dimension of Spurious Features

We assume that the dimension of spurious features $z_s$ is smaller than the dimension of the core features $z_c$, which is denoted as

$$\begin{aligned} &dim(z_s) = p\, dim(z),\ dim(z_c) = (1-p)\, dim(z) \\ &s.t.\ p \in (0, 0.5) \end{aligned} \tag{S1}$$

We set $p = 0.25$ in all of the datasets we evaluated.

[1] https://github.com/ecreager/eiil

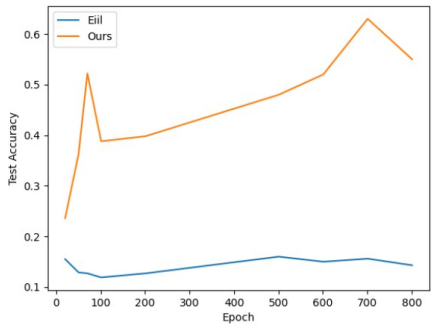


**Figure S1:** Test accuracy of EIIL and our method with different epochs on the Colored MNIST[II] dataset.

### B.5 Early Stopping in Reference Models

We notice the impact of epochs on the reference model. EIIL trains the reference model at a low epoch perhaps because it is proposed that ERM can capture good spurious features early in the training [4]. However, Figure S1 reports that ERM used as a reference model in EIIL with the Colored MNIST[II] dataset still suffers from a pitfall in invariant learning due to the low test accuracy. Even at about 20 epochs, which is much less than the epochs in invariant learning, ERM still can not capture effective spurious features perhaps because the diversity of bias-conflicting samples limits the inductive bias of ERM. We find that our method outperforms EIIL which uses ERM as a reference model and it shows that our method captures better spurious features with less impact on the inductive bias of early stopping. The shortcoming of ERM used as a reference model is that we can not obtain more spurious features by adjusting the epochs in the training step. If ERM does not work well with early stopping, which means it can not capture sufficient spurious features early in the training, we have only a few methods to make the environment inference a better performance unless we can obtain more heterogeneous information.

## C Theoretical Details

### C.1 Metric: Spurious Feature Score

Given a test environment $\mathcal{E}_{te} := \{x_i, y_i\}_{i=1}^{n_{test}}$, and supposed that each $x_i$ is correspond to $K$ kinds of core features, each sample $x$ is the concatenation of spurious feature $x^s$ and core feature $x^c$. $\mathbb{I}(\cdot)$ is the indication function.

$$\frac{1}{n^{test}} \frac{1}{K} \sum_{i=1}^{n^{test}} \sum_{j=1}^{K} \mathbb{I}(f([x_i^s, x_i^c]) = f([x_i^s, x_j^c])) \tag{S2}$$

### C.2 Metric: Invariant Penalty

$$\max_{q} \left\| \nabla_{\overline{w}} \widetilde{R}^e(\overline{w} \cdot \Phi, q) \right\|^2, \forall e \in \{\mathcal{E}_0, \mathcal{E}_1\} \tag{S3}$$

### C.3 Proof Details of Identifiability

We show the proof details of the identifiability of disentangled representations in our VAE model. First, we detail the definition of identifiability.

**Definition 1 (Identifiability)** *In VAE model, we say $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ is identifiable if*

$$p_\theta(x) = p_{\tilde{\theta}}(x) \tag{S4}$$

*We can recover observed $x$ by estimated distribution $p_{\tilde{\theta}}$ from independent factors $\{z_1, z_2, ..., z_n\}$ if the solutions of which is unique.*

VAE model is not identifiable if there is no valid prior [3], which means there are multiple solutions of the latent variable by given observed $x$. Non-uniqueness means we may obtain useless latent variables for downstream tasks because we can not know which solution can truly recover the observed $x$, and in other words, the solution may be unexplainable. Previous researchs [1, 2] about nonlinear ICA have shown that Eq. S5 leads to identifiability, when a conditionally factorized prior distribution over the latent variables in VAE model, denoted as $p_\theta(z|u)$.

$$p_\theta(x, z|u) = p_f(x|z)p_{T,\lambda}(z|u) \tag{S5}$$

where $u$ is denoted as an auxiliary variable, which can be considered as a prior over VAE model, and $\theta = (f, T, \lambda)$ is the parameters of VAE model.

We assume that $z$ is the concatenation of spurious features $z_s$ and core features $z_c$, which is denoted as $z = [z_s, z_c]$, and $z_s$ is independent of $z_c$ at a high probability. Then Eq. S6 holds.

$$p(z_c|z_s) = 1 \tag{S6}$$

We assign the $z_s$ to the core features w.r.t class labels when training the model, and we can approximately replace $z_s$ to auxiliary variables $u$ because labels are considered as the conditioned prior. Our VAE decoder distribution can be reformulated as

$$\begin{aligned} p_\theta(x|z) &= p_\theta(x|z_s, z_c) \\ &= p_\theta(x|u, z_c) \\ &= p_\theta(x|u, z_c)p(z_c|u) \\ &= p_\theta(x, z_c|u) \end{aligned} \tag{S7}$$

Then the core features $z_c$ can be identifiable when spurious features $z_s$ are applied with core feature labels. It can be questioned why not directly apply the core feature labels to the entire latent variable $z$, but only a certain part of $z$. If we do that, VAE model can have the same pitfall as ERM under distributional shift because spurious features are still entangled with core features in latent variables and both kinds of features can not be identifiable for classification prediction.

# References


[1] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[2] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[3] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[4] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.