LEARNING TASK-RELEVANT FEATURES VIA CONTRASTIVE INPUT MORPHING

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental challenge in artificial intelligence is learning useful representations of data that yield good performance on a downstream classification task, without overfitting to spurious input features. Extracting *task-relevant* predictive information becomes particularly challenging for high-dimensional, noisy, real-world data. We propose Contrastive Input Morphing (CIM), a representation learning framework that learns input-space transformations of the data to mitigate the effect of irrelevant input features on downstream performance via a triplet loss. Empirically, we demonstrate the efficacy of our approach on various tasks which typically suffer from the presence of spurious correlations, and show that CIM improves the performance of other representation learning methods such as variational information bottleneck (VIB) when used in conjunction.

1 INTRODUCTION

At the heart of modern machine learning is the problem of representation learning, or extracting features from raw data that enable predictions with high accuracy (Hinton & Salakhutdinov, 2006; Vincent et al., 2010; Chen et al., 2016; Van Den Oord et al., 2017; Oord et al., 2018). Despite the recent successes of deep neural networks (Dean et al., 2012; LeCun et al., 2015), their rapidly growing size and large-scale training procedures, coupled with high-dimensional data sources, pose significant challenges in learning models that perform well on a given task without overfitting to spurious input features (Zhang et al., 2016; Ilyas et al., 2019; Geirhos et al., 2020). As a result, trained networks have been shown to fail spectacularly on out-of-domain generalization tasks (Beery et al., 2018; Rosenfeld et al., 2018) and for rare subgroups present in data (Hashimoto et al., 2018; Goel et al., 2020), among others.

A wide range of methods have been proposed to tackle this problem, including regularization, data augmentation, leveraging causal explanations, and self-training (Srivastava et al., 2014; Chen et al., 2020b; Sagawa et al., 2019; Chen et al., 2020b). In particular, prior art places a heavy emphasis on lossless access to the input data during training, then constructing a high-level representation which extracts the necessary information. Yet it is reasonable to assume that in some cases, we desire access to only a *subset* of the input which is relevant to the task – for example, the background color in an image of a "7" is unnecessary for identifying its digit class. The fundamental challenge, then, is discerning which parts of the input are relevant without requiring privileged information (e.g., the nature of the downstream task) at training time.

Our approach, Contrastive Input Morphing (CIM), uses labeled supervision to *learn input-space transformations of the data* that mitigate the effect of irrelevant input features on predictive performance. Though the Data Processing Inequality (Cover, 1999) states that no amount of input processing can increase its mutual information (MI) with the predictive variable, we propose to transform the data in such a way that it makes it easier for the model to extract the *relevant* predictive information for the downstream task – that is, we attempt to increase the amount of usable information for our representations (Xu et al., 2020). We emphasize that our method does not assume access to the exact nature of the downstream task, such as attribute labels for rare subgroups.

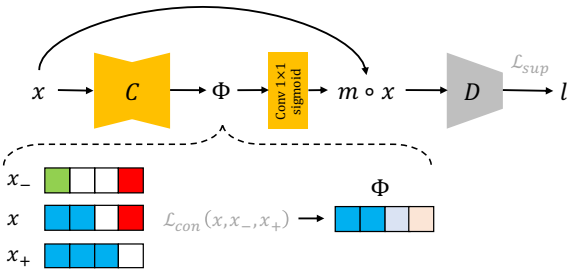


Figure 1: An end-to-end flowchart for the CIM training procedure. C refers to the TN while D refers to the classifier (discriminator). $\mathcal{L}_{con}$ forces the learned input transformation Φ to upweight the task-relevant blue features which are present in both x and x_+, and to downweight the spurious red feature which is shared by x and x_-.

The key workhorse of CIM is an auxiliary network called the Transformation Network (TN). Leveraging ideas from neural style transfer (Gatys et al., 2015; Li et al., 2017b), the TN is trained via a triplet loss on feature correlation matrices (Schroff et al., 2015; Koch, 2015). Intuitively, this objective uses the shared information from competing classes ("negative examples") as a proxy for spurious correlations, while leveraging the shared information within the same class ("positive examples") as a heuristic for task-relevancy (Khosla et al., 2020). The framework for CIM is quite general: it is (1) complementary to MI-based representation learning techniques such as variational information bottleneck (VIB) (Alemi et al., 2016); and (2) can be used as a plug-in module for training any classifier. For the flowchart of the training procedure of the CIM refer to Figure 1.

Empirically, we evaluate CIM on three settings that suffer from spurious correlations: classification with nuisance background information, out-of-domain (OOD) generalization, and improving accuracy uniformly across subgroups. In the first task, CIM outperforms ERM on colored MNIST and improves over the ResNet-50 baseline on the Background Challenge (Xiao et al., 2020). Similarly, CIM outperforms relevant baselines using ResNet-18 on the VLCS dataset (Torralba & Efros, 2011) for OOD generalization. For subgroup accuracies, CIM outperforms both supervised and unsupervised methods on CelebA (Liu et al., 2015) in terms of worst-group accuracy (by 1.7% and 41.4% respectively), while outperforming unsupervised methods by up to 12.9% on Waterbirds.

In summary, our contributions in this work can be outlined as follows:

1. We propose CIM, a method demonstrating that lossy access to input data helps extract good task-relevant representations.
2. We show that CIM is complementary to existing methods, as the learned transformations can be leveraged by other MI-based representation learning techniques such as VIB.
3. We empirically verify the robustness of the learned representations to spurious correlations on a variety of tasks (Section 4).

2 Preliminaries

We consider the standard supervised learning setup where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is the input variable, and $y \in \mathcal{Y} = \{1, \ldots, k\}$ is the set of corresponding labels. We assume access to samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ drawn from an underlying (unknown) joint distribution $p_{\text{data}}(x, y)$, and use capital letters to denote random variables, e.g. X and Y. We use $P(X, Y)$ to denote their joint distribution as well as $P(\cdot)$ for the respective marginal (e.g. $P(X)$ for the marginal distribution of X).

2.1 Background and Problem Setup

Our goal is to learn a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $f_\theta \in \Theta$ achieves low error according to some loss function $\ell : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$. Specifically, we minimize the empirical risk:

$$\mathcal{L}_{\text{sup}}(\theta) = \mathbb{E}_{x,y \sim p_{\text{data}}(x,y)}[\ell(f_\theta(x), y)] \approx \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) \tag{1}$$

In addition to good classification performance, we aim to learn representations of the data, which: (a) are highly predictive of the downstream task; and (b) do not rely on spurious input features. That is, the learned representations should be *task-relevant*.

Information bottleneck. A natural way to measure "task-relevance" in random variables is to consider the total amount of information that a compressed (stochastic) representation Z contains about the input variable X and the output variable Y. In particular, information bottleneck (IB) (Tishby et al., 2000; Chechik et al., 2005; Alemi et al., 2016) is a framework which utilizes mutual information (MI) to quantify this dependence via the following objective:

$$\min_{P(Z|X)} I(X;Z) - \beta I(Z;Y) \tag{2}$$

where $\beta > 0$ controls the importance of obtaining good performance on the downstream task. Given two random variables X and Y, $I(X;Y)$ is computed as $D_{\text{KL}}(P(X,Y)||P(X)P(Y))$, where D_{KL} denotes the Kullback-Leibler (KL) divergence between two probability distributions.

The IB framework can be extended to account for additional sources of input data that is known to contain irrelevant information about the predictive task. This setting, known as IB with side information (Chechik & Tishby, 2003), adds a term in the IB objective, which simultaneously minimizes the MI between this nuisance variable and the learned representation. Concretely, given random variables (X, Y_+, Y_-) where Y_+ denotes the task of interest and Y_- denotes a spurious auxiliary variable, the objective becomes:

$$\min_{P(Z|X)} I(X;Z) - \beta(I(Z;Y_+) - \gamma I(Z;Y_-)) \tag{3}$$

where γ is another tunable hyperparameter for the nuisance task. We note that this framework bears resemblance to triplet-based losses such as (Schroff et al., 2015; Koch, 2015), as well as contrastive learning approaches that leverage MI maximization (Linsker, 1988; Hjelm et al., 2018; Oord et al., 2018; Tian et al., 2019; Khosla et al., 2020). It is also in line with the InfoMin principle suggested by (Tian et al., 2020), for learning good views in self-supervised contrastive learning.

Although the MI framework is compelling, as it captures arbitrarily complex dependencies between random variables, there exist several challenges with their use in practice. The difficulty of computing MI in high dimensions, for example, is well-documented, demands the use of various neural estimators (Barber & Agakov, 2003; Gutmann & Hyvärinen, 2010; Oord et al., 2018; Belghazi et al., 2018; Poole et al., 2019; Song & Ermon, 2019). Additionally, approaches such as IB posit restrictive assumptions on the relationships between Y_+ and Y_-; namely, that they must be conditionally independent given X, which is difficult to enforce.

3 Contrastive Input Morphing

Motivated by the above challenges, we propose to approximate the information content between task-relevant and -irrelevant features via correlations in higher-dimensional feature spaces. This procedure helps our method *learn* the appropriate input-space transformations.

3.1 Measuring Relevance via Correlations

Another way to measure "task-relevance" in random variables is to consider their conditional dependencies, as captured by their covariances. Specifically, consider a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ that takes in an input and returns a representation $\phi(x)$ that is of the same dimensionality as the original input x. We can use this feature map to construct a covariance (Gram) matrix of $\phi(x)$, where $\Sigma_{XX} = \phi(x)^T\phi(x)$. Although the covariance only measures linear dependencies between the input, we can capture more complex relationships via an arbitrarily complex feature map ϕ.

Training Procedure: For the Transformation Network (TN), we utilize a convolutional autoencoder to obtain a reconstructed image of the same dimensionality as the input, as shown in Figure 2. Our method operates over triplets: (x, x_+, x_-), where (x, x_+) denote examples from the same class while x_- is an example from a different class than x. We use a supervised contrastive loss to train the network, similar in spirit to (Khosla et al., 2020).

Specifically, we learn an intermediate feature map $\phi : \mathcal{X} \to \mathbb{R}^{(H\times W)\times C}$ using the TN that takes in an input x and returns a representation $\phi(x)$ that is of the same dimensionality as the original

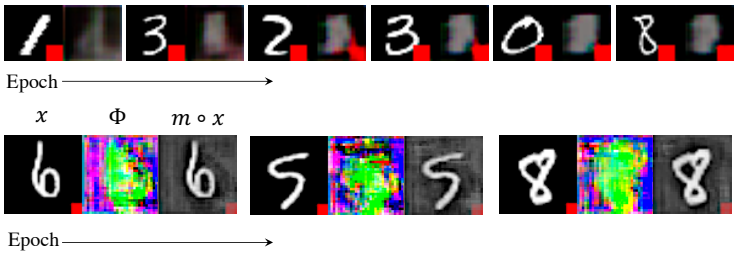


Figure 2: Captured similarities on MNIST. **Top:** When training the TN as an autoencoder, the triplet loss forces the network to reconstruct the shared spurious features between examples. **Bottom:** When training CIM for classification, the input transformation Φ highlights the task-relevant digit while de-emphasizing the uninformative sources of variation (the background and red square).

input, where $(H \times W)$ denotes the height and width of the image, and C denotes the number of channels. We use this feature map to construct a Gram matrix of the input features, where $\Sigma_{XX} = \phi(x)^T\phi(x)$. Then, the triplet loss encourages the positive examples' Gram matrix representations to move closer together in embedding space to those of the input, while ensuring that the negative examples' representations are further apart:

$$\mathcal{L}_{\text{con}}(\phi) = \min_{\phi} ||\Sigma_{XX}, \Sigma_{X_+X_+}||^2 - \max(\alpha, ||\Sigma_{XX}, \Sigma_{X_-X_-}||^2) \tag{4}$$

for some margin $\alpha > 0$. The output of $\phi(\cdot)$ is then passed through a 1×1 2-D convolution layer with a sigmoid activation to produce a (single channel) soft "mask" $m(x)$, which is then multiplied with the original input image x to obtain the final representation $\psi(x) = x \circ m(x)$. Finally, the classifier $f_\theta(\cdot)$is trained on the transformed input image $\psi(x)$.

Learning Objective: The overall loss function can be written as:

$$\mathcal{L}(\phi, \theta) = \lambda\mathcal{L}_{\text{con}}(\phi) + \mathcal{L}_{\text{sup}}(\theta) \tag{5}$$

where λ is a multiplier which controls the contribution of the TN loss from $\mathcal{L}_{\text{con}}(\phi)$ from Equation 4 and $\mathcal{L}_{\text{sup}}(\theta)$ is the standard cross entropy loss for multi-class classification. The parameters for the transformation network (ϕ) and the classifier (θ) are trained jointly. In our experiments, we found that values of $\lambda = 0.0001$ worked well.

It is well known that $\mathcal{L}_{\text{con}}$ can be interpreted as minimizing a specific form of Maximum Mean Discrepancy (MMD) (Gatys et al., 2015; Li et al., 2017b). For the identity map $\phi(\cdot)$, Equation 5 is equivalent to minimizing MMD between two kernelized inputs where the specific kernel is the second-order polynomial kernel. In this way, CIM's Transformation Network can also be seen as minimizing the distance between the mean embeddings of the underlying distributions for X and X_+ while simultaneously maximizing the distance for those of X and X_-.

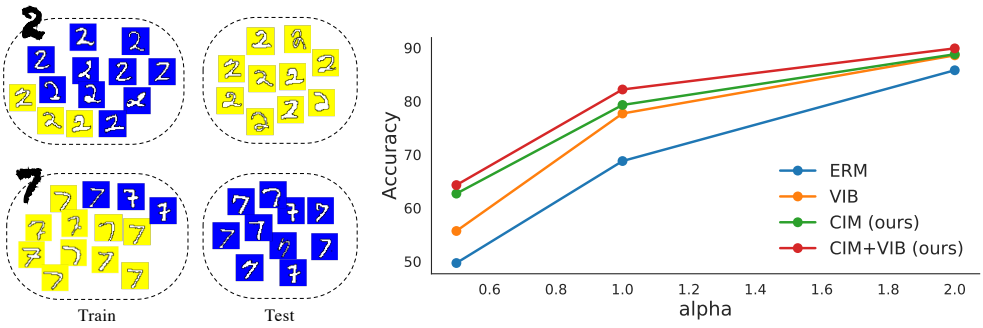
3.2 A Motivating Example

We present a concrete example for the intuition behind our approach using the MNIST dataset (LeCun, 1998). We first construct a challenging input reconstruction task, in which a red square is placed on the bottom right of all samples and the model is trained to reconstruct a random digit that is different from the input's class. As the TN is trained as an independent autoencoding module, we find that the TN learns to pick up shared signals across inputs (i.e., the black background) before converging to the red square as the source of shared (spurious) correlations among examples.

Next, we evaluate whether we can remove this source of variation for digit classification by passing a lossy version of the input into the classifier (Figure 1). As shown in Figure 2 (bottom), the input transformation learned by the model (i.e. $\psi(x) = m \circ x$) de-emphasizes the shared features while highlighting the task-relevant features.

4 Experimental Results

For our experiments, we are interested in empirically investigating the following questions:



(a) Train vs. test distributions

(b) Classification results; higher is better.

Figure 3: Results with nuisance background information for Colored MNIST. In (a), the train and test sets are constructed such that a classifier will achieve low accuracy by relying on background color. In (b), CIM and CIM + VIB outperform relative baselines on digit classification.

1. Are CIM's learned representations robust to spurious correlations in the input features?
2. Does the input transformation learned by CIM improve domain generalization?
3. How well can CIM preserve classification accuracy across subgroups?

Datasets: We consider various datasets to test the effectiveness of our method. We first construct a colored variant of MNIST to demonstrate that CIM successfully ignores nuisance background information in a digit classification task, then further explore this finding on the Background Challenge (Xiao et al., 2020). Next, we evaluate CIM on the VLCS dataset (Torralba & Efros, 2011) to demonstrate that the input transformations help in learning representations that generalize to out-of-domain distributions. Then, we study two benchmark datasets, CelebA (Liu et al., 2015) and Waterbirds (Wah et al., 2011; Zhou et al., 2017; Sagawa et al., 2019), to show that CIM preserves subgroup accuracies.

Models: We use different classifier architectures depending on the downstream task. While ResNet-50 is the default choice for most datasets, we also utilize Inception-ResNetV2 (Szegedy et al., 2016) to obtain better performance, ResNet-18 for a fair comparison with existing OOD generalization techniques, and PointNet (Qi et al., 2017) for 3D point cloud classification. We also experiment with Variational Information Bottleneck (VIB) (Alemi et al., 2016) as both a complementary and competing approach to CIM, and use ResNet-50 as the VIB encoder. We refer the reader to Appendix A.2 for additional details on model architectures and hyperparameters. We note that the transformed inputs and the feature maps Φ are semantically meaningful as shown in Figure 4.

4.1 Classification with Nuisance Background Information

Colored MNIST: First, we assess whether CIM can distinguish between two MNIST digit classes (2 and 7) in the presence of a spurious input feature (background color). As outlined in Figure 3(a), we construct a dataset such that a classifier will achieve low accuracy by relying on background color. For a given proportion α, we color $\alpha\%$ of all digits labeled "2" in the training set with blue backgrounds, and color the remaining $(1-\alpha)\%$ labeled "7" with yellow backgrounds. We vary this proportion by $\alpha = \{0.5\%, 1\%, 2\%\}$. At test time, we color all the digits labeled "2" in blue, while coloring the "7" digits in yellow. As shown in Figure 3 (b), CIM is better able to utilize relevant information for the downstream classification task in comparison to ERM by 13%, 10.5%, and 3% on models trained with $\alpha = \{0.5\%, 1\%, 2\%\}$ respectively. Perhaps more interestingly, a hybrid approach of VIB + CIM outperforms all other methods – this suggests that the input transformations learned by CIM are indeed preserving task-relevant information that can be better leveraged by InfoMax methods such as VIB. More experimental details can be found in Appendix A.2.

The Background Challenge: Next, we evaluate whether the favorable results from MNIST translate to a more challenging setup, and test CIM on the Background Challenge (Xiao et al., 2020). The Background Challenge is a public dataset consisting of ImageNet-9 (Deng et al., 2009) test sets with varying amounts of foreground and background signals, designed to measure the extent to which

	Original (↑)	Mixed-same (↑)	Mixed-rand (↑)	BG-gap (↓)
Baseline (Xiao et al., 2020)	96.3	89.9	75.6	14.3
CIM (Ours)	97.1	**90.4**	79.7	10.7
CIM + VIB (Ours)	**97.5**	90.3	**79.9**	**10.4**

Table 1: Results from the Background Challenge on ImageNet-9 using ResNet-50. Our method outperforms the relevant baselines across all three datasets. The difference between Mixed-same and Mixed-rand is referred to as the background gap (BG-gap), which indicates average robustness to varying backgrounds from different image sources.

deep classifiers rely on spurious features for image classification. As shown in Table 1, CIM outperforms the original ResNet-50's performance by 4.1% on Mixed-rand, 0.8% on Mixed-same, and 0.5% on the original test set. Mixed-rand refers to the setting where the foreground is overlaid onto a random background, while Mixed-same corresponds to the test set where the foreground is placed on a background from the same class. These results demonstrate that CIM indeed learns task-relevant representations without relying on nuisance background information.

4.2 CIM Generalizes Over Different Domains

In this experiment, we evaluate CIM on OOD generalization performance using the VLCS benchmark (Torralba & Efros, 2011). VLCS consists of images from five object categories shared by the PASCAL VOC 2007, LabelMe, Caltech, and Sun datasets, which are considered to be four separate domains. We follow the standard evaluation strategy used in (Carlucci et al., 2019), where we partition each domain into a train (70%) and test set (30%) by random selection from the overall dataset. As summarized in Table 2, CIM outperforms state-of-the-art methods based on ResNet-18 on each domain, bolstering our claim that using a lossy transformation of the input is helpful for learning task-relevant representations that generalize across domains.

Method	Caltech	LabelMe	Pascal	Sun	Average
DeepC (Li et al., 2018b)	87.47	62.06	64.93	61.51	68.89
CIDDG (Li et al., 2018b)	88.83	63.06	64.38	62.10	69.59
CCSA (Motiian et al., 2017)	92.30	62.10	67.10	59.10	70.15
SLRC (Ding & Fu, 2017)	92.76	62.34	65.25	63.54	70.15
TF (Li et al., 2017a)	93.63	63.49	69.99	61.32	72.11
MMD-AAE (Li et al., 2018a)	94.40	62.60	67.70	64.40	72.28
D-SAM (D'Innocente & Caputo, 2018)	91.75	57.95	58.59	60.84	67.03
JiGen (Carlucci et al., 2019)	96.93	60.90	70.62	64.30	73.19
Asadi et al. (Asadi et al., 2019)	98.11	63.61	74.33	67.11	75.79
CIM (Ours)	**98.63**	**66.67**	**75.36**	**69.62**	**77.57**

Table 2: Multi-source domain generalization results (%) on the VLCS dataset with ResNet-18 as the base network. All reported numbers are averaged over three runs.

4.3 CIM Preserves Subgroup Performance

In this experiment, we investigate whether representations learned by CIM perform well on all subgroups on the CelebA and Waterbirds datasets. Preserving good subgroup-level accuracy is challenging for naive ERM-based methods, given their tendency to latch onto spurious correlations (Kim et al., 2019; Arjovsky et al., 2019; Sagawa et al., 2020; Chen et al., 2020b). Most prior works leverage privileged information such as group labels to mitigate this effect (Ben-Tal et al., 2013; Vapnik & Izmailov, 2015; Sagawa et al., 2019; Goel et al., 2020; Xiao et al., 2020). As TN in CIM is trained to capture task-relevant features and minimize nuisance correlations between classes, we hypothesize that CIM should perform well at the subgroup level *even without explicit group label information*.

For a fair comparison with the prior work, we use ResNet-50 as the backbone classifier for the CIM , but also train both ERM and CIM with an Inception-ResNetV2 (Szegedy et al., 2016) backbone to

assess the impact of using a larger model (denoted by ERM* and CIM*, respectively). We also use ResNet-50 for VIB's encoder and InfoMask's discriminator (see Appendix A.2). Table 3 shows that CIM outperforms both supervised and unsupervised methods on CelebA in terms of worst-group accuracy (2.4% improvement over CAMEL, the top-performing supervised model), and outperforms unsupervised models while significantly improving over ERM on the Waterbirds dataset (16.7% increase). We emphasize that the favorable performance of CIM is obtained *without using subgroup labels*, in contrast with previous approaches. We refer the reader to Appendix B.3 for further details and ablation studies regarding the different components of our method.

Dataset	Method	Unsupervised (group-level)	Worst-group acc.	Overall acc.
CelebA	GDRO (Sagawa et al., 2019)	✗	82.2	90.9
	GDRO* (Sagawa et al., 2019)	✗	49.4	91.4
	CAMEL (Goel et al., 2020)	✗	**83.5**	92.9
	ERM (Vapnik, 2013)	✓	41.4	91.4
	ERM* (Vapnik, 2013)	✓	35.9	91.1
	VIB (Alemi et al., 2016)	✓	40.6	90.5
	InfoMask (Taghanaki et al., 2019)	✓	43.8	88.4
	CIM (Ours)	✓	85.2	91.1
	CIM + VIB (Ours)	✓	**85.9**	90.2
	CIM* (Ours)	✓	85.1	93.8
Waterbirds	GDRO (Sagawa et al., 2019)	✗	83.8	89.4
	GDRO* (Sagawa et al., 2019)	✗	75.2	97.4
	CAMEL (Goel et al., 2020)	✗	**89.6**	90.9
	ERM (Vapnik, 2013)	✓	59.7	95.4
	ERM* (Vapnik, 2013)	✓	54.8	95.8
	VIB (Alemi et al., 2016)	✓	69.9	95.3
	InfoMask (Taghanaki et al., 2019)	✓	58.4	94.9
	CIM (Ours)	✓	72.6	94.8
	CIM + VIB (Ours)	✓	76.4	95.4
	CIM* (Ours)	✓	**77.9**	96.4

Table 3: Average and worst-group accuracies for CelebA and Waterbird benchmark datasets. Methods without group-level supervision (i.e. with ✓) are preferable. * refers to methods with Inception-ResNetV2 backbone instead of ResNet-50. CIM outperforms both supervised and unsupervised methods on the CelebA dataset as well as unsupervised methods on the Waterbirds dataset. It also achieves favorable performance relative to supervised methods on Waterbirds.

5 Related Work

Our work bridges several lines of work in contrastive learning and learning representations that are robust to spurious correlations.

Contrastive representation learning. There has been a flurry of recent work in contrastive methods for representation learning, which encourages an encoder network to map "positive" examples closer together in a latent embedding space while spreading the "negative" examples further apart (Oord et al., 2018; Hjelm et al., 2018; Wu et al., 2018; Tian et al., 2019; Arora et al., 2019; Chen et al., 2020a). Included are triplet-based losses (Schroff et al., 2015; Koch, 2015) and noise contrastive estimation losses (Gutmann & Hyvärinen, 2010). In particular, recent work (Tian et al., 2020; Wu et al., 2020) has shown that *minimizing* MI between views while maximizing predictive information of the representations with respect to the downstream task, leads to performance improvements, similar to IB (Chechik & Tishby, 2003). While most contrastive approaches are self-supervised, (Khosla et al., 2020) utilizes class labels as part of their learning procedure, similar to our approach. We emphasize that CIM is not meant to be directly comparable to the aforementioned techniques, as our objective is to learn input transformations of the data that are task-relevant.

(a) Transformed inputs on 3D point clouds.

(b) Sampled Φ on various datasets.

Figure 4: Qualitative visualizations of the learned representations from CIM. (a) Morphed point cloud objects of the Modelnet40 (Wu et al., 2015) dataset. The first row shows the raw input, while the second row shows the morphed input. The first two columns are samples from the `flowerpot` category while the next three are from the `sink` class. (b) Samples of learned Φ. Left to right: CelebA, Waterbirds, and the Background Challenge datasets.

Robustness of representations Several works have considered the problem of learning relevant features that do not rely on spurious correlations with the predictive task (Heinze-Deml & Meinshausen, 2017; Sagawa et al., 2020; Chen et al., 2020b). Though (Wang et al., 2019) is similar in spirit to CIM, they utilize gray-level co-occurrence matrices as the spurious (textural) information of the input images, then regress out this information from the trained classifier's output layer. Our method does not solely rely on textural features and can learn any transformation of the input space that is relevant for the downstream task of interest. Although CIM also bears resemblance to InfoMask (Taghanaki et al., 2019), our method is not limited to attention maps. (Kim et al., 2019) uses an MI-based objective to minimize the effect of spurious features, while (Pensia et al., 2020) additionally incorporates regularization via Fisher information to enforce robustness of the features. On the other hand, CIM uses an orthogonal approach to learn robust representations via higher-order correlations in the features.

Information in representations There is a rich body of work which focuses on quantifying the amount of information necessary to perform well on a downstream task (Achille & Soatto, 2018). CIM is reminiscent of InfoMax (Linsker, 1988) and IB-based approaches (Tishby et al., 2000; Alemi et al., 2016) which propose to maximize the MI in the learned representations with the predictive random variables. In particular, (Chechik & Tishby, 2003; Chechik et al., 2005; Goyal et al., 2020) is most similar to our setup where they consider additional (nuisance) predictive information. Rather than using MI, we draw inspiration from the style transfer literature (Gatys et al., 2015; Li et al., 2017b; Krichene et al., 2018; Sastry & Oore, 2019) to compare correlations between feature activations of relevant versus irrelevant examples during training.

6 Conclusion

In summary, we considered the problem of extracting representations with *task-relevant* information from high-dimensional data. We introduced a new framework, CIM, which learns input-space transformations of the data via a triplet loss to mitigate the effect of irrelevant input features on downstream performance. Through experiments on (1) classification with nuisance background information; (2) OOD domain generalization; and (3) preservation of uniform subgroup accuracy, we showed that CIM achieves good performance despite the presence of spurious correlations in the data and outperforms most relevant baselines. Additionally, we demonstrated that CIM is complementary to other representation learning frameworks such as VIB. For future work, it would be interesting to test different types of distance metrics for the triplet loss, to explore whether CIM can be used as an effective way to *learn* views for unsupervised contrastive learning, and to investigate label-free approaches for learning the input transformations.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Nader Asadi, Mehrdad Hosseinzadeh, and Mahdi Eftekhari. Towards shape biased unsupervised representation learning for domain generalization. *arXiv preprint arXiv:1909.08245*, 2019.

David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Gal Chechik and Naftali Tishby. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems*, pp. 881–888, 2003.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of machine learning research*, 6(Jan):165–188, 2005.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *arXiv preprint arXiv:2006.10032*, 2020b.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pp. 1223–1231, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017.

Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, pp. 187–198. Springer, 2018.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

Anirudh Goyal, Yoshua Bengio, Matthew Botvinick, and Sergey Levine. The variational bandwidth bottleneck: Stochastic evaluation on an information budget. *arXiv preprint arXiv:2004.11935*, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9012–9020, 2019.

Gregory Koch. Siamese neural networks for one-shot image recognition. 2015.

Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang, Xinyang Yi, Lichan Hong, Ed Chi, and John Anderson. Efficient training on very large corpora via gramian estimation. *arXiv preprint arXiv:1807.07187*, 2018.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018a.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018b.

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

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

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

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *arXiv preprint arXiv:2005.04345*, 2020.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv preprint arXiv:1912.12510*, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

Saeid Asgari Taghanaki, Mohammad Havaei, Tess Berthier, Francis Dutil, Lisa Di Jorio, Ghassan Hamarneh, and Yoshua Bengio. Infomask: Masked variational latent representation to localize chest disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 739–747. Springer, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 ARCHITECTURES

In Figure 5, we show the detailed TN architectures used for RGB and point-cloud data.

```
[ Conv2D(32, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  MaxPooling2D((2, 2), padding='same'),
  Conv2D(16, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  MaxPooling2D((2, 2), padding='same'),
  Conv2D(8, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  MaxPooling2D((2, 2), padding='same'),
  Conv2D(8, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  UpSampling2D((2, 2)),
  Conv2D(16, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  UpSampling2D((2, 2)),
  Conv2D(32, (3, 3), activation='relu', padding='same'),
  BatchNormalization(),
  UpSampling2D((2, 2)),
  Conv2D(3, (1, 1), padding='same')
  BatchNormalization(name='Φ') ]
```

a) RGB input

```
[ Conv1D(32, 1, activation='relu'),
  BatchNormalization(),
  Conv1D(64, 1, activation='relu'),
  BatchNormalization(),
  Conv1D(128, 1, activation='relu'),
  BatchNormalization(),
  Conv1D(64, 1, activation='relu'),
  BatchNormalization(),
  Conv1D(32, 1, activation='relu'),
  BatchNormalization(),
  Conv1D(3, 1, activation='tanh'),
  BatchNormalization(name='Φ') ]
```

b) Point-cloud input

Figure 5: Architecture of the TN for both RGB and point-cloud inputs.

A.2 HYPERPARAMETER CONFIGURATIONS AND TRAINING DETAILS

Variational Information Bottleneck (VIB). We used ResNet-50 as the encoder in VIB because most methods we compare CIM with are based on ResNet-50. We tested two different settings for VIB after the encoder: (a) apply KL regularization on encoder's last layer L_f of size (1, 2048) and compute the cross-entropy loss on the regularized feature vector; (b) apply KL on the feature vector similar to (a), but add 3 fully connected layers of `(1024, ReLu, batch normalization)`, `(512, ReLu, batch normalization)`, and `(256, ReLu)`, then calculate the cross-entropy loss; (c) add a fully connected layer of size 512 after L_f, then follow the steps as in (a). For colored MNIST we used architecture (c) and trained the model using Adam optimizer with a learning rate set to 0.0001 and batch size of 64. For celebA and Waterbirds, we used architecture (b) with Adam optimizer and learning rate of 0.001 and batch size of 64. For all the above experiments we set the weight for KL regularization term to be 0.001 and the standard deviation of ϵ to be 0.1.

InfoMask. We used the default architecture (Taghanaki et al., 2019) except for changing the encoding part to be ResNet-50. For celebA experiments, we used Adam optimizer with a learning rate of 0.0001 and a batch size of 32. For Waterbirds, we trained the model using SGD optimizer with a learning rate of 0.001 and a momentum of 0.9. Similar to VIB, we set the KL term weight to be 0.001 and the standard deviation of ϵ to be 0.1. We tested different threshold values for the masking function and obtained the best results with just soft masking i.e. when the threshold is set to zero.

Point Cloud Experiments. For PointNet, we used Adam optimizer with a learning rate of 0.0001 and a batch size of 32. We trained both the original and CIM based model with rotated and jittered input data.

Colored MNIST. We resized images to $(64 \times 64 \times 3)$ and trained all the models using Adam optimizer with a learning rate of 0.0001 and batch size of 64. For VIB, we set the KL divergence contribution weight to 0.001.

Domain Generalization. We use ResNet-18 as the backbone to make a fair comparison with state-of-the-art. We train CIM using Adam optimizer with learning rate of 0.0001 and batch size of 64. We use the same training and test splits as those used in the work with (Carlucci et al., 2019).

For CIM-based models, we set $\lambda = 0.0001$ and other hyper-parameters are summarized in Table 4. To control the level of input re-weighting, we minimize negative entropy on m with a Lagrangian multiplier $\zeta = 0.00001$.

Table 4: Hyper-parameters for our CIM and CIM* methods.

Task	Method	Optimizer	Batch size	Input size
MNIST	CIM	`Adam(lr=0.0001)`	64	(64, 64, 3)
CelebA	CIM	`Adam(lr=0.0001)`	64	(224, 224, 3)
	CIM*	`Adam(lr=0.0001)`	64	(224, 224, 3)
Waterbirds	CIM	`SGD(lr=0.001, momentum=0.9)`	32	(224, 224, 3)
	CIM*	`SGD(lr=0.001, momentum=0.9)`	64	(224, 224, 3)
Background challenege	CIM	`SGD(lr=0.001, momentum=0.9)`	64	(224, 224, 3)
	CIM*	`SGD(lr=0.001, momentum=0.9)`	64	(224, 224, 3)
Point-clouds	CIM	`Adam (lr=0.0001)`	32	(2048, 3)

B Additional Experimental Results

B.1 Background Challenge

We include for completeness the entirety of the results from (Xiao et al., 2020). We note that our results are not directly comparable with those from other architectures (e.g. WRN-50x2), as we used ResNet-50 as our base classifier.

	`Original` (↑)	`Mixed-same` (↑)	`Mixed-rand` (↑)	`BG-gap` (↓)
AlexNet	86.7	76.2	54.2	22.0
ShuffleNet	95.7	86.7	69.4	17.3
VGG16-BN	97.6	91.0	78.0	13.0
WRN-50x2	97.2	90.6	78.0	12.6
ResNet-50	96.3	89.9	75.6	14.3
CIM (Ours)	97.1	90.4	79.7	10.7
CIM + VIB (Ours)	97.5	90.3	79.9	10.4

Table 5: Results from the Background Challenge on ImageNet-9 using ResNet-50. The difference between `Mixed-same` and `Mixed-rand` is referred to as the background gap (`BG-gap`), which indicates average robustness to varying backgrounds from different image sources.

B.2 3D Point Cloud Classification

In Table 6, we report the classification results on normal and rotated objects. As the first row of the table summarizes, PointNet performs well on average on the 40 classes. However, when we increase spurious correlations by rotating the objects, class-wise accuracies significantly drop, resulting in a 16.1% performance degradation in the average accuracy of the model (second row). After applying CIM, the spurious correlation between different categories is reduced, thus class-wise accuracy of challenging objects is improved (third row).

Table 6: Modelnet40 (Wu et al., 2015) point cloud classification results.

Method	`flowerpot`	`radio`	`sink`	`xbox`	`dresser`	Avg. 40 classes
PointNet (Qi et al., 2017)	15	60	70	75	83	88.8
PointNet (Qi et al., 2017)	5	35	45	50	60	72.7
CIM [Ours]	**25**	**45**	**60**	**65**	**75**	**73.2**

B.3 ABLATION STUDIES

We construct an ablation study on the CelebA dataset to study the effects of the Gramian-based contrastive loss. As shown in Table 7, we find that learning a simple attention-like weighting matrix without any regularization performs better than ERM. We also observed that having both positive and negative samples in the TN's loss function performs better compared to having only positives or negatives. It is worth mentioning the negative samples have a greater impact on the performance in comparison to the positivies.

Table 7: Ablation study of CIM on the CelebA dataset. Rand x_- corresponds to the common contrastive learning strategy i.e. using random negative samples and augmented version of the input as positives. Only m refers to applying m on the input without any contrastive regularization. Only x_+ and Only x_- refer to leveraging only positive or negative terms in our contrastive Gramian loss, respectively. The highlighted column shows the worst group accuracy.

Method	Blonde male	Blonde female	Non-blonde female	Non-blonde male	Average acc.
Rand x_-	69.53	96.05	90.36	96.85	93.33
Only m	78.13	98.07	86.79	95.31	91.35
Only x_+	80.47	97.49	87.75	95.22	91.72
Only x_-	82.81	97.99	85.83	93.62	90.27
CIM	85.16	97.94	86.46	93.74	90.63