Improving Novel Anomaly Detection by Learning Domain-Invariant Representations in Latent Space

Padmaksha Roy[1 [0000−0002−9571−1117]](✉), Ming Jin[1 [0000−0001−7909−4545]], Himanshu Singhal[1 [0000−0002−0474−8126]], Tyler Cody[1 [0000−0001−9215−5816]], and Kevin Choi[2 [0009−0004−1818−2401]]

[1] Virginia Tech, Blacksburg, Virginia, USA
{padmaksha,jinming,himanshusinghal,tcody}@vt.edu
[2] AI Center of Excellence, Deloitte, Mclean, Virginia, USA kevchoi@deloitte.com

Abstract. Zero-day anomaly detection is critical in industrial applications where novel, unforeseen threats can compromise system integrity and safety. Traditional detection systems often fail to identify these unseen anomalies due to their reliance on in-distribution data. Domain generalization addresses this gap by leveraging knowledge from multiple known domains to detect out-of-distribution events. In this work, we introduce a multi-task representation learning technique that fuses information across related domains into a unified latent space. By jointly optimizing classification, reconstruction, and mutual information regularization losses, our method learns a minimal(bottleneck), domain-invariant representation that discards spurious correlations. This latent space decorrelation enhances generalization, enabling the detection of anomalies in unseen domains. Our experimental results demonstrate significant improvements in zero-day or novel anomaly detection across diverse anomaly detection datasets.

Keywords: Representation Learning · OOD Detection · Multi-task Learning.

1 Introduction

Anomaly detection is a fundamental task in various applications, enabling the early identification of unusual patterns in network traffic, system logs, or user behavior that may signal intrusions or malicious activities [23, 27]. As cyber threats evolve and novel attacks—such as zero-day vulnerabilities—emerge, traditional defenses often fall short, leading to severe disruptions and data breaches. In many real-world applications, training and test data stem from different distributions, making out-of-distribution (OOD) generalization a critical challenge. Standard deep neural networks excel when the training and testing data are drawn from the same distribution; however, their performance degrades when confronted with unseen domains. Existing approaches such as few-shot learning and meta-learning [19, 25, 17, 16, 21] attempt to bridge this gap but often

require target domain data during training or otherwise risk embedding biases from specific domains. Our approach addresses these challenges by targeting a latent space that embodies a minimal sufficient representation for the downstream task of OOD classification. We consider a scenario where the samples from different domains or datasets have distinct feature correlation structures. High-dimensional data poses unique challenges due to the curse of dimensionality. In such spaces, conventional distance measures lose their discriminative power because the relative contrast between the nearest and farthest neighbors diminishes—a phenomenon highlighted by the principle of concentration of distance. Inspired by the principle of relevant information(PRI) preservation [35], we design a latent space classification loss that aims to regularize the latent space by minimizing the mutual information content between the input and latent space, effectively decorrelating class-specific feature correlation information of the original data. To guarantee that the latent space preserves sufficient input information, we incorporate a reconstruction loss that compels the model to accurately reconstruct the input data from its latent embedding. This prevents over-compression and ensures that the latent space retains the necessary structure for the task. These two losses guide the cross-entropy loss to preserve only the relevant information required for accurate classification. Multi-task learning facilitates learning representations from multiple diverse domains and the joint optimization help improve generalization to unseen domains. By integrating these objectives, our framework works as a zero-shot multi-task learning system. We mix data from multiple source domains with cross-domain samples and also attempt to decorrelate dataset specific spurious correlation information with the mutual information (MI) penalty. This strategy ensures that the learned latent space is invariant to domain-specific correlation information, thereby enhancing generalization to unseen OOD classes without requiring any target domain data during training. Our main contributions can be summarized as follows:

- We propose a novel classification framework that leverages mutual information regularization and reconstruction loss to guide the latent space toward retaining only the most relevant features for out-of-distribution (OOD) classification. The result is a *compressed, invariant* representation that effectively discards *spurious* domain-specific information.
- We demonstrate that integrating data from multiple sources and cross-domains with varying *correlation* patterns enhances *coverage*, improving generalization to unseen domains.
- Our domain-invariant latent space analysis mitigates the adverse effects of *high-dimensionality*. Experiments demonstrate an 8%-15% increase in average precision, and recall and a 4%-9% improvement in average AUC-ROC across all source/IN, cross-domain, and OOD datasets.

2 Related Work

Domain generalization techniques can be grouped into the following primary categories: domain invariant representation learning, meta-learning,latent di-

mension regularization, and metric learning. *1) Domain Invariant Representation Learning:* This method aims to identify domain invariant representations that can be extended to unseen domains. The crux of these strategies, as seen in works such as [33], is to filter out domain-specific insights while maintaining cross-domain information. Notable studies employing autoencoders, such as [1], amalgamate multiple domains during training, augmented by data enhancement techniques, to extract domain-invariant characteristics. These features then demonstrate superior generalization to out-of-distribution data. Another study, Maximum Mean Discrepancy Adversarial Autoencoder (MMD-AAE) [24], in the context of few-shot learning, emphasizes aligning varied domain distributions to a generic prior distribution while engaging in adversarial feature learning. An innovative approach is suggested in [14], where a domain-centric masking technique is applied to learn both domain-specific and domain-invariant features. This will facilitate efficient source domain classification and sufficient generalization to target domains. In [15], a noise-enhanced supervised autoencoder reconstructs and classifies both inputs and their reconstructions, using intra-class correlation to show improved feature discrimination and generalization. Moreover, the authors [29] propose domain generalization through domain-invariant representation that uniformly distributes across multiple source domains. Their approach employs moment alignment of distributions and enforces feature disentanglement via an entropy loss. The DIFEX [18] paper employs knowledge distillation to capture internally-invariant Fourier phase features and aligns cross-domain correlations to extract mutually-invariant representations.

2) Meta-learning: This approach employs learning from several related tasks for domain generalization, as observed in works such as [21, 28, 32]. The study in [30] introduces a technique to discern a domain interdependent projection leading to a latent space. This space minimizes biases in the data while preserving the inherent relationship across multiple domains. Model Agnostic Meta-Learning (MAML) has also been extended to latent dimension settings by performing the gradient-based adaptation in the low dimensional space instead of the higher dimensional space of model parameters [31]. Zero-shot learning [9] aims at learning models from seen classes and inferring on samples whose categories were unseen during the training process.

3) Information Bottleneck Principle and Metric Learning: In contrast to the aforementioned methodologies, our strategy propels direct disentanglement or decorrelation between multiple training domains. An information-theoretic perspective on variance-invariance-covariance has been provided here [3] in the context of self-supervised learning which helps to achieve generalization guarantees for downstream supervised learning tasks. Adversarial learning-based domain adaptation methods are prone to negative transfer which hurts the generalization performance [4]. Metric learning aims to learn a representation function that can map higher-dimensional data to a latent embedded space. The authors [5] propose mixing target labels with training samples to improve the quality of representations or embeddings for classification purposes.

4) Other Related Works: The authors [6] suggest using the statistics of softmax outputs to estimate both the probability of error and the likelihood of a test sample being out-of-domain. They compare the performance of this approach by directly using the raw softmax output probabilities as a measure of confidence. The paper [7] addresses the problem of domain shift when a learned model tends to degrade heavily on a target domain via unsupervised domain adaptation by learning a common feature map from multiple source domains by minimizing the domain distribution discrepancy between those multiple source domains. The authors in [10] use nearest-neighbor distance for flexible OOD detection without strict assumptions, while [11] address spurious correlations by developing causal tools to distinguish invariant features, thereby improving generalization.

Our approach is inspired by the principle of relevant information (PRI) [35], aiming to learn a compressed latent space that retains only the relevant information for downstream tasks. By combining data from multiple domains and de-correlating their spurious correlations, we encourage the network to learn invariant representations. This multi-task representation learning method ensures that the latent space captures minimal, sufficient information for classification while discarding irrelevant, domain-specific details.

3 Problem Formulation

In our domain generalization problem, let $\mathcal{C} = \{0, 1, \ldots, K\}$ denote the complete set of class labels, which we partition into three disjoint subsets; $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_c \cup \mathcal{C}_o, \quad \mathcal{C}_s \cap \mathcal{C}_c = \mathcal{C}_s \cap \mathcal{C}_o = \mathcal{C}_c \cap \mathcal{C}_o = \emptyset$. For example, if $\mathcal{C}_s = \{1, 3, 5\}$ (source domain), then $\mathcal{C}_c = \{2, 4, 6\}$ (cross-domains) and $\mathcal{C}_o = \{9, 10, 11\}$ (OOD). During training, we have access only to samples from the source and cross-domains. Formally, the training set is defined as $\mathcal{S}_{\text{train}} = \bigcup_{i=1}^{M} \{(x_j^i, y_j^i) \mid y_j^i \in \mathcal{C}_s \cup \mathcal{C}_c,\ j = 1, \ldots, N_i\}$, where M is the total number of tasks and N_i is the number of samples in task i. Our objective is to learn a model that generalizes to unseen OOD classes $y \in \mathcal{C}_o$ by leveraging multi-task representation learning. We enforce domain-invariant feature extraction through joint optimization over classification, reconstruction, and mutual information regularization losses, thereby encouraging a disentangled latent space that tends to forget spurious correlations. The extension to multiple domains necessitates the definition of a multi-task learning objective over all the M source and cross domains which can be given as

$$\mathcal{L}_{rec}\left(S_{train}; \theta, \phi\right) = \sum_{i=1}^{M} \left\| f_\theta^{(i)} \left(g_\phi^{(i)} \left(X^i \right) \right) - X^i \right\|_2^2 \tag{1}$$

In this expression, $g_\phi^{(i)}$ and $f_\theta^{(i)}$ denote the encoder and decoder functions respectively for each of the M sources and cross domains, X^i is the input training data from a particular domain. We aggregate the reconstruction loss across different source and cross-domain datasets, ensuring that the total loss accounts for all input domains. Basically, the reconstruction error is computed separately for each domain and then summed to form the overall reconstruction loss.

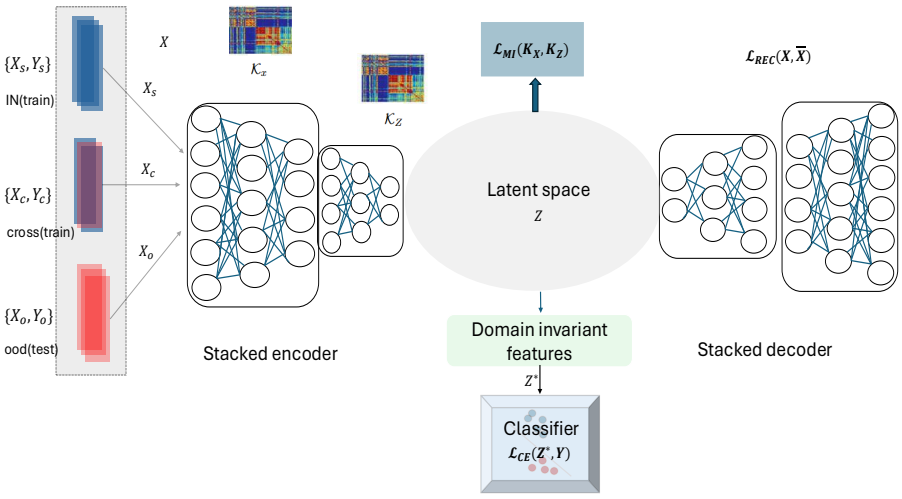


Fig. 1: Training the Multi-task Latent Space Regularized Encoder-Decoder Model (MTLS-RED). During testing, the trained latent space is directly used to classify new samples.

3.1 Mutual Invariance Regularization

In information theory, the dependence measure or the total correlation between the feature variables is measured as the statistical independence in each dimension and is expressed as the Kullback Leibler(KL) divergence between the joint probability distribution and the marginal distribution of the features [26]. We enforce de-correlation between the input and the latent kernel space—spanning multiple source and cross-domains—by introducing a mutual information minimization penalty that explicitly reduces dependencies between input and latent space kernels in the form of decorrelation. The matrix-based Renyi's second-order entropy [26] of a normalized positive definite(NPD) matrix $\mathcal{K}_x$, estimated on $l \times l$ samples in the input space, where l is the batch size, can be given as

$$\hat{H}_2(\mathcal{K}_x) = \frac{1}{1-\alpha} \log_2 \left(\sum_{k=1}^{l} \lambda_k(\mathcal{K}_x)^\alpha \right), \tag{2}$$

where the Gram matrix $\mathcal{K}_x$ is obtained by evaluating the positive definite (PSD) kernel on all l pairs of training samples in a batch of training data, that is, and $\lambda_k(X)$ denotes the k^{th} eigenvalue of the input kernel matrix $\mathcal{K}_x$ of the l_{th} batch, Here, $\alpha = 2$ considering Renyi's second-order entropy.

Similarly, Renyi's quadratic entropy of the latent space kernel $\mathcal{K}_Z$ of size $l \times l$ is estimated as

$$\hat{H}_2(\mathcal{K}_z) = \frac{1}{1-\alpha} \log_2 \left(\sum_{k=1}^{l} \lambda_k(\mathcal{K}_z)^\alpha \right), \tag{3}$$

The argument in equation (3) is called the information potential. In the above section, we use the matrix-based second-order Renyi's entropy ($\alpha = 2$) [26] to evaluate the entropy or the uncertainty of the latent and the input space in terms of the normalized eigenspectrum of the Hermitian matrix of the projected data in the Hilbert space. Now, we can estimate the matrix-based second-order joint entropy between the latent space kernel Z and the input space kernel X as

$$\hat{H}_2\left(\mathcal{K}_x, \mathcal{K}_z\right) = H_2\left(\frac{\mathcal{K}_x \circ \mathcal{K}_z}{tr\left(\mathcal{K}_x \circ \mathcal{K}_z\right)}\right), \tag{4}$$

where $\circ$ represents the Hadamard product. Based on the above definitions, we calculate the joint entropy of the latent and the input space with the help of the matrix-based normalized Renyi's entropy of the latent space and the input space kernels. The joint entropy is used to derive the mutual information between the input and the latent space.

The Mutual Information Divergence We use the matrix-based mutual information divergence to estimate the mutual information between the latent and input space kernels. Minimizing the mutual information indirectly results in decorrelating the feature correlation that exists in the original input space which helps in improving the generalization performance. The mutual information during each batch of the training can be estimated as

$$\hat{MI}(\mathcal{K}_x; \mathcal{K}_z) = \hat{H}_2\left(\mathcal{K}_x\right) + \hat{H}_2\left(\mathcal{K}_z\right) - \hat{H}_2\left(\mathcal{K}_x, \mathcal{K}_z\right), \tag{5}$$

where $\hat{H}_2\left(\mathcal{K}_X, \mathcal{K}_Z\right)$, is the second-order joint entropy between the latent and the input kernel space. Minimizing this divergence as a regularization penalty in the final loss objective will aid in preserving useful disentangled information in the latent space during each iteration of the training process.

3.2 The Multi-Task Learning Objective

In our latent space multi-task learning approach, we leverage the label information of the multiple source and cross-domain encoded data in the latent space during the training process. In our approach, we do a joint optimization of the classification and the reconstruction loss along with the mutual information penalty in the latent space. The total loss calculated over all the M tasks can be written as

$$\begin{aligned}\mathcal{L}\left(\mathcal{S}_{train}, Z; \phi, \theta, \sigma\right) = \min_{\phi,\theta,\sigma} \sum_{i=1}^{M} \Big\{ & \mathcal{L}_{ce}\left(g_\phi\left(X^i\right), y^i\right) \\ & + \beta \cdot \mathcal{L}_{MI}\left(X^i; Z^i, \sigma\right) + \lambda \cdot \mathcal{L}_{rec}\left(X^i; \phi, \theta\right) \Big\},\end{aligned} \tag{6}$$

where, $\mathcal{L}_{ce}$ is the cross-entropy loss calculated on the latent space encoding considering the binary classification problem, given as,

$$\mathcal{L}_{ce}\left(g_\phi\left(X^i\right), y^i\right) = -\left(y^i \log\left(\mathcal{S}_y^i\left(g_\phi\left(X^i\right)\right)\right) + \left(1 - y^i\right) \log\left(1 - \left(\mathcal{S}_y\left(g_\phi\left(X^i\right)\right)\right)\right)\right)$$

Algorithm 1 The Multi-task Latent Space Regularized Encoder-Decoder Model (MTLS-RED)

Input:
Source domain data $\{X_{s1}, X_{s2}, ..., X_{sm}\}$, $X_s \in \mathbf{R}^d$, $\forall m \in \{1, 2, 3, ..\}$
Cross-domain data $\{X_{c1}, X_{c2}, ..., X_{cn}\}$, $X_c \in \mathbf{R}^d$, $\forall n \in \{4, 5, 6, .., \}$
Out-of-distribution (OOD) datasets $\{X_{o_1}, X_{o_2}, ..., X_{o_k}\}$, $X_o \in \mathbf{R}^d$, $\forall k \in \{7, 8, 9, .., \}$ (used for testing only)
Source and cross-domain labels $\{y_i^m\}_{i=1}^n$, $\forall m \in \{1, 2, 3, 4, 5, 6, ..., M\}$
Initialize encoder (E) and decoder (D) weights:
$\mathbf{W}_\phi \in \mathbf{R}^{d_x \times d_z}$, $\mathbf{W}_\theta \in \mathbf{R}^{d_z \times d_x}$
Initialize kernel bandwidths: σ_x, σ_y (learnable)
Set learning rates $\alpha_1, \alpha_2, \alpha_\sigma$
while not end of epochs **do**:
  for batch = 1 to total batches N **do**:
    Sample mini-batch data $\{X_i\}_1^l \in \mathbf{R}^d$, where l is batch-size
    Compute RBF kernels for input space $\mathcal{K}_{x_l}$ and latent space $\mathcal{K}_{z_l}$ of size $l \times l$
    Compute mutual information between input space X_l and latent space Z_l using matrix-based Rényi's entropy:

$$MI\left(\mathcal{K}_{X_l}; \mathcal{K}_{Z_l}\right)$$

    Perform a forward pass on encoder $E\left(X_{\phi_i}\right)$
    Compute total batch loss:

$$\mathcal{L}_l = \mathcal{L}_{ce}\left(X^l, y^l\right) + \lambda \mathcal{L}_{rec}\left(X^l, X^{l'}\right) + \beta \mathcal{L}_{\mathcal{MI}}\left(X^l || Z^l\right)$$

    Update $\mathbf{W}_\phi$, $\mathbf{W}_\theta$, and σ_x, σ_z
    $\mathbf{W}_{\phi_{t+1}} \leftarrow \mathbf{W}_{\phi_t} - \alpha_1 \nabla_\phi \mathcal{L}_l\left(\theta, \phi, \sigma\right)$
    $\mathbf{W}_{\theta_{t+1}} \leftarrow \mathbf{W}_{\theta_t} - \alpha_2 \nabla_\theta \mathcal{L}_l\left(\theta, \phi, \sigma\right)$
    $\sigma_{x_{t+1}}, \sigma_{z_{t+1}} \leftarrow \sigma_{x_t}, \sigma_{z_t} - \alpha_\sigma \nabla_\sigma \mathcal{L}_l\left(\theta, \phi, \sigma\right)$
  end for
end while
Output: Trained MTLS-RED model with optimized encoder-decoder weights $\mathbf{W}_\phi, \mathbf{W}_\theta$ and learned kernel bandwidths σ_x, σ_y

$\mathcal{L}_{MI}$ is the disentanglement or de-correlation loss between the latent space and the input space expressed in the form of mutual information divergence measured in their kernel space, given in eq: 5, $\mathcal{S}_y$ is the softmax function applied on the encoded data $g_\phi(x)$, $\mathcal{L}_{rec}$ is the reconstruction loss, ϕ, θ are the encoder and decoder parameters. The σ represents the kernel bandwidth, a crucial parameter for estimating mutual information between the input and latent space.

We guide the cross-entropy loss by incorporating mutual information regularization between the latent and input spaces. This regularization discourages the retention of irrelevant information in the latent representation, with its strength governed by the hyperparameter β The parameter β regulates the trade-off between reducing dependencies in the latent space and maintaining classification performance. During joint optimization, we aim to balance the reconstruction

loss and mutual information regularization. The parameter λ controls the reconstruction weight, determining the extent of compression we want to enforce in the latent space.

4 Experiments

In this section, we demonstrate the performance of our proposed model on benchmark cybersecurity and healthcare datasets.

4.1 Dataset

- **CSE-CIC-IDS2018 [20]** This is a publicly available cybersecurity dataset that is made available by the Canadian Cybersecurity Institute (CIC). It consists of 7 major kinds of intrusion datasets. We use SOLARIS, GOLDENEYE as source domain data, INFILTRATION, BOTNET as cross-domain data, and RARE, SLOWHTTPS, HOIC and a BENIGN dataset of a different day as the OOD test classes.
- **CICIoT 2023 [20]** This is a state-of-the-art dataset for profiling, behavioral analysis, and vulnerability testing of different IoT devices with different protocols from the network traffic, consisting of 7 major attack classes. We use BENIGN, DoS, and DDoS as source data, RECON, as cross-domain data and WEB, MIRAI as OOD test data.
- **CICIoMT 2024 [20]** This is a benchmark dataset to enable the development and evaluation of Internet of Medical Things (IoMT) security solutions. The attacks are categorized into five classes. We use BENIGN, DDoS, DoS as source-domain, RECON, and SPOOFING as cross-domain, and MQTT as OOD data.
- **Arrythmia** This dataset is about atrial fibrillation (also called AFib or AF) which is a quivering or irregular heartbeat (arrhythmia) that can lead to blood clots, stroke, heart failure, and other heart-related complications. The dataset contains five classes/categories: N (Normal), S (Supraventricular ectopic beat), V (Ventricular ectopic beat), F (Fusion beat), and Q (Unknown beat).

4.2 Baselines

We consider the following models related to multi-task representation learning and few-shot learning as baselines.

- **Correlation Alignment for Deep Domain Adaptation (CORAL)** [34] This work has been employed for supervised domain adaptation, aligning source and target covariances to enhance OOD generalization.
- **Multi-task Autoencoder (MTAE)** [1] This encoder-decoder model optimizes reconstruction error across multiple domains in a supervised manner, jointly training sources and cross-domain data with label information in a two-stage process.

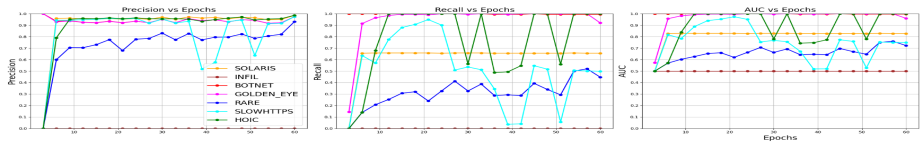
- **Minimum Mean Discrepancy-Autoencoder(MMD-AE)**[24, 2] This paper uses the MMD measure as regularization for domain generalization between multiple cross-domain data. We use it as a few-shot learning method where the cross-domain data are added to improve the OOD generalization.
- **Noise Enhanced Supervised Autoencoder (NSAE)** [15]This model jointly predicts input labels, reconstructs inputs as noisy samples, and refines them through an additional fine-tuning step using a supervised classifier.
- **Domain-invariant Feature Exploration for Domain Generalization (DIFEX)** [18] This paper utilizes mutual invariance to extract cross-domain features for OOD classification, capturing domain-specific semantics through internal invariance while preserving shared information, and extends CORAL with an additional regularization term.

4.3 Training Strategy

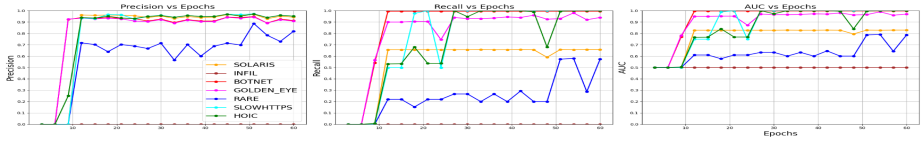
To achieve robust generalization, we arbitrarily categorize the datasets into three groups: source domain datasets, cross-domain datasets, and out-of-distribution (OOD) datasets. The OOD datasets are reserved exclusively for testing purposes, serving as an evaluation benchmark for assessing model generalization. Our training strategy focuses on enhancing OOD performance by leveraging source domain data to improve learning on cross-domain datasets. To accomplish this, we systematically mix different proportions of source domain data with cross-domain data, integrating them into the benign dataset to construct the final training set. Additionally, we experiment with different combinations of source and cross-domain datasets to identify the most effective configurations for improving coverage across all three dataset categories—source, cross-domain, and OOD. Our training strategy is detailed in MTLS-RED Algorithm 1.

Selecting the cross-domains, source and OOD domains In [36], the authors argue that learning a model that generalizes to unseen data can be facilitated when the *covariance* (or correlation structure) among features is well-conditioned and sufficiently diverse. In other words, if the training data exhibit meaningful variations or "patterns of dependencies" across features, then a function that captures those variations can more reliably extrapolate beyond the training distribution. Hence, by adding source domain data (with one correlation structure) to cross-domain data (with a different correlation structure), we produce a more *varied* training distribution—one that exposes the learner to multiple ways in which features can co-vary. The learner, in turn, is incentivized to find a representation that extracts the stable, non-spurious relationships across these distributions.

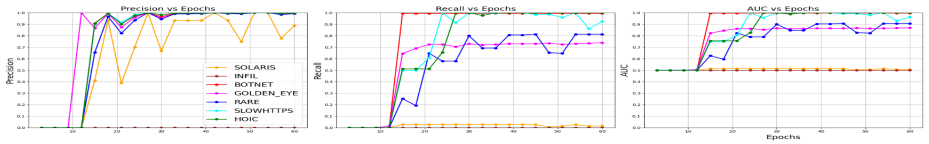
This approach is inspired by the paper's emphasis on the role of well-conditioned covariances for successful extrapolation, suggesting that diverse training correlations expand the set of feature configurations on which the model is trained, thus boosting generalization performance in truly novel test domains. DOS and DDOS share similar feature correlations, while MIRAI and WEB differ, and



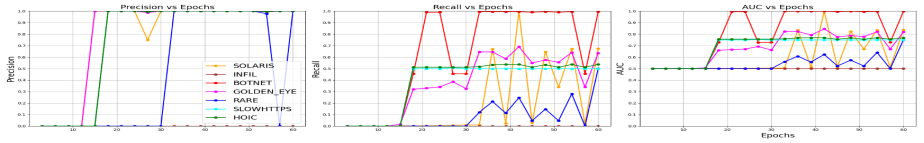
(a)



(b)



(c)



(d)

Fig. 2: Precision, recall, and accuracy plots for the rarest class (RARE, in blue), which has only 525 samples in the CIC-IDS dataset using training data from GOLDENEYE (source) and BOTNET (cross) domains. Figures (a) show precision, recall, and AUC over epochs without regularization on validation data; (b) apply MI = 0.01, reconstruction = 0.99; (c) apply MI = 0.99, reconstruction = 0.01, (d)use equal weights of 0.5. High MI regularization (case (c)) leads to over 10-20% improvement and stability across all metrics. Higher MI penalty helps in achieving better classification of the RARE class

Table 1: We report **accuracy** (with standard deviation) of the proposed and baseline methods on the CIC-CSE-IDS dataset, where cross-domain data is gradually added to the source domain during training in the range (0–50%). Best test accuracies for each model are highlighted. The OOD domains are used only for test/evaluation purposes. During train, each anomaly dataset has equal amount (50%) of BENIGN samples added to it, i.e, the train and test datasets are balanced (equal normal and anomaly samples).

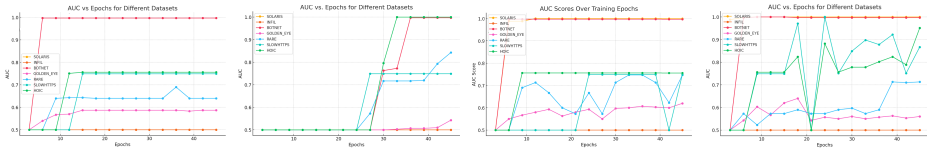
		SOURCE DOMAIN		**CROSS DOMAIN**		**OOD DOMAIN**			
Model	**Percent**	SOLARIS	GOLDEYE	INFIL	BOTNET	RARE	HOIC	HTTPS	BENIGN
MTAE	0%	**99.97 (2.5)**	92.70 (1.3)	04.00 (0.1)	09.50 (0.2)	00.30 (0.0)	00.38 (0.0)	02.30 (0.1)	97.98 (2.4)
	20%	99.50 (2.5)	**99.70 (2.5)**	**69.40 (2.5)**	**79.50 (2.5)**	61.30 (1.5)	27.38 (0.7)	32.30 (0.8)	**98.11 (2.5)**
	30%	81.60 (2.0)	79.60 (2.0)	69.70 (1.7)	78.50 (2.0)	65.40 (1.6)	**50.00 (1.3)**	42.40 (1.1)	62.00 (1.6)
	50%	61.61 (1.5)	61.30 (1.5)	31.30 (0.8)	61.60 (1.5)	**72.60 (1.8)**	48.50 (1.2)	**69.80 (1.7)**	83.32 (2.1)
MMD-AE	0%	**99.99 (2.5)**	**99.29 (2.5)**	00.53 (0.0)	**99.98 (2.5)**	**55.47 (1.4)**	**99.98 (2.5)**	**99.69 (2.5)**	**99.71 (2.5)**
	20%	99.98 (2.5)	92.55 (0.1)	**22.55 (0.6)**	99.83 (2.5)	12.19 (0.3)	41.34 (1.0)	98.77 (2.5)	98.67 (2.5)
	30%	99.78 (2.5)	02.52 (0.1)	15.35 (0.4)	99.83 (2.5)	12.19 (0.3)	41.27 (1.0)	99.69 (2.5)	98.61 (2.5)
	50%	99.96 (2.5)	01.88 (0.1)	12.30 (0.3)	99.83 (2.5)	14.13 (0.4)	41.12 (1.0)	56.21 (1.4)	99.14 (2.5)
NSAE	0%	99.99 (2.5)	99.98 (2.5)	00.09 (0.0)	99.99 (2.5)	**77.03 (1.9)**	**99.90 (2.5)**	**97.02 (2.4)**	99.58 (2.5)
	20%	99.80 (2.5)	**99.99 (2.5)**	04.82 (0.1)	34.92 (0.9)	36.04 (0.9)	00.00 (0.0)	05.16 (0.1)	**99.80 (2.5)**
	30%	**99.99 (2.5)**	03.16 (0.1)	00.32 (0.0)	**99.83 (2.5)**	59.36 (1.5)	15.90 (0.4)	00.35 (0.0)	99.69 (2.5)
	50%	99.98 (2.5)	34.36 (0.9)	**53.66 (1.3)**	99.83 (2.5)	12.36 (0.3)	00.00 (0.0)	19.33 (0.5)	99.40 (2.5)
CORAL	0%	59.18 (1.5)	90.61 (2.3)	30.45 (0.8)	00.80 (0.0)	08.40 (0.2)	55.06 (1.4)	03.40 (0.1)	69.94 (1.7)
	20%	61.53 (1.5)	12.64 (0.3)	**31.79 (0.8)**	**50.43** (1.3)	**33.74 (0.9)**	41.31 (1.0)	**50.38** (1.3)	67.54 (1.7)
	30%	38.85 (1.0)	**99.99 (2.5)**	0.00 (0.0)	0.01 (0.0)	22.96 (0.6)	82.21 (2.1)	0.00 (0.0)	95.01 (2.4)
	50%	**99.99 (2.5)**	38.85 (1.0)	00.01 (0.0)	00.00 (0.0)	22.96 (0.6)	**82.21 (2.1)**	00.01 (0.0)	**99.19 (2.5)**
MTLS-RED	0%	**98.83 (2.2)**	96.08 (2.2)	72.41 (2.1)	**96.12 (2.4)**	28.01 (0.7)	80.99 (2.0)	73.93 (1.8)	73.23 (1.8)
	20%	79.00 (2.0)	78.85 (2.0)	78.71 (2.0)	78.68 (2.0)	78.88 (2.0)	79.13 (2.0)	78.82 (2.0)	79.01 (2.0)
	30%	83.90 (2.1)	80.70 (2.0)	70.70 (2.0)	83.80 (2.1)	76.50 (1.9)	81.96 (2.0)	**85.00 (2.1)**	77.10 (1.9)
	50%	86.46 (2.0)	**89.33 (2.0)**	**79.20** (2.0)	89.12 (2.0)	**78.99 (2.0)**	**89.33** (2.0)	79.25 (2.0)	**79.64 (2.0)**

GOLDEN and SOLARIS exhibit distinct correlations from INFIL and BOTNET. We aim to enhance generalization by training on datasets with varying feature correlation structures while ensuring overlapping marginal distributions for effective extrapolation. [3]

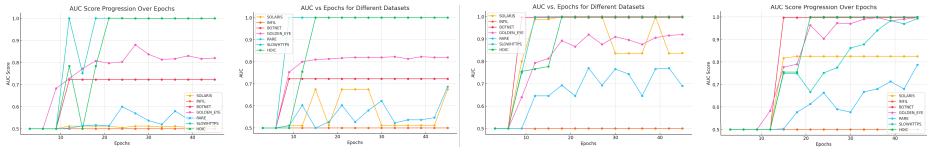
4.4 Hyperparameter Sensitivity

In the joint optimization framework, achieving an optimal balance between the compression regularization (λ) and the mutual information regularization (β) is crucial for ensuring strong generalization across all classes. The reconstruction loss, weighted by λ, governs the degree of compression in the latent space—excessive compression may lead to the loss of essential features, while insufficient compression can result in overfitting to the input distribution. Meanwhile, the mutual information regularization, controlled by β, acts as a de-correlation penalty, reducing redundant dependencies between the latent and input spaces. Properly tuning β ensures that the latent representation retains only the most discriminative information for classification. Our findings indicate

[3] https://github.com/padmaksha18/MTRAE/blob/main/mtrae/mtl-reg-cse-cic-ids-V333333-noisy-equal-cross.ipynb



(a) BOTNET (0.0, 0.0; 0.6760) (b) BOTNET (0.5, 0.5; 0.7332) (c) BOTNET, SOLARIS (0.1,0.1; 0.7670) (d) BOTNET, SOLARIS(0.01, 0.99; 0.7983)



(e) GOLDEN-EYE (0.0, 0.0; 0.749) (f) GOLDEN-EYE (0.01, 0.99; 0.7721) (g) GOLDEN-EYE, BOTNET (0.1, 0.1; 0.849) (h) GOLDEN-EYE, BOTNET (0.1, 0.9; 0.8915)

Fig. 3: From left to right, the plots show improved average AUC-ROC as datasets are combined and regularization is applied, enhancing generalization to unseen domains. We evaluate this using seven CIC-CSE-IDS attack datasets with equal benign samples, reporting results as (reconstruction weight, MI penalty, Average AUC on all datasets).

that prioritizing entropy regularization (higher β) while reducing the emphasis on reconstruction loss (lower λ) yields the best overall model performance across diverse scenarios, reinforcing the importance of controlled compression and structured disentanglement in the latent space.

Importance of kernel bandwidth In a non-parametric estimation method such as our mutual information penalty, the kernel bandwidth σ plays a crucial role. By learning σ jointly with the encoder and decoder, we adapt the kernel scale to match the data distribution's complexity. We vary the proportion of cross-domain data in training, ranging from 0%-50% of the source data, to analyze the effects of the de-correlation penalty and reconstruction regularization under different scenarios. In Tables 1 and 2, we observe that as the proportion of cross-domain data in training increases, the performance of most baseline models deteriorates on the IN distribution. In particular, OOD domain data remain completely unseen throughout the training process. Table 1 reveals an intriguing trend: as cross-domain data increases to $40\% - 50\%$, adjusting the kernel bandwidth and hyperparameters β and λ allows us to train a model that achieves comprehensive generalization across all training and test datasets. Figure 3 illustrates the improvement in average AUC-ROC as datasets are combined and regularization is introduced. Both strategies—dataset combination and reg-

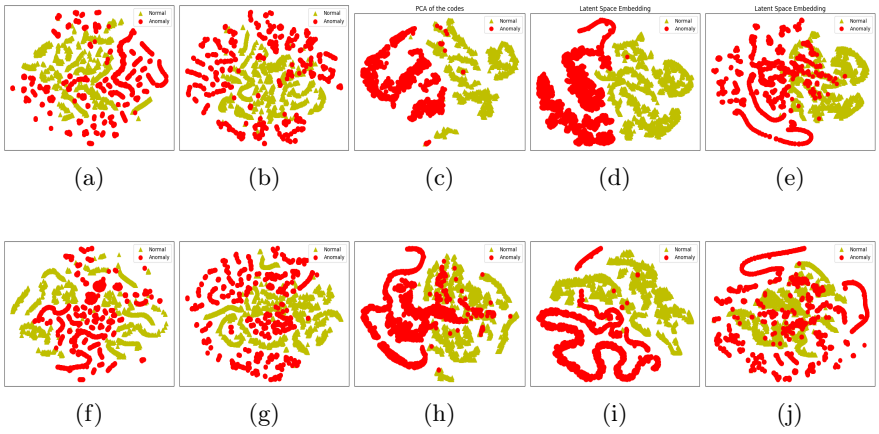


Fig. 4: T-SNE projection of the latent space of without regularization case (bottom row) and MTL-RED (top row) for some of the attacks in CIC-IDS and CIC-IOMT/IOT: SOLARIS, RARE, DOS, DDOS, and RECONAISSANCE. Subfigures (a)–(e) correspond to MTL-RED, and (f)–(j) to no regularization case.

Table 2: We report **accuracy** (with std deviation) of the proposed and baseline methods on the CIC-IOMT/IOT dataset Other details are similar as Table 1.

Model	**Percent**	**SOURCE DOMAINS**		**CROSS DOMAINS**		**OOD DOMAINS**			
		DDOS	DOS	RECON	SPOOF	MQTT	MIRAI	WEB	BENIGN
MTAE	0%	99.96 (2.5)	99.99 (2.6)	54.07 (1.5)	46.98 (1.3)	78.43 (2.1)	**99.97 (2.8)**	30.55 (0.8)	97.53 (2.8)
	20%	99.99 (2.8)	99.99 (2.7)	98.66 (2.3)	71.84 (1.4)	78.43 (2.0)	79.38 (1.8)	**30.55 (0.7)**	97.53 (2.2)
	30%	99.99 (2.4)	99.99 (2.5)	**99.99 (2.5)**	**75.55 (1.9)**	**99.93 (2.9)**	80.07 (2.3)	21.38 (0.6)	98.63 (2.3)
	50%	**99.99 (2.4)**	**99.99 (2.7)**	98.53 (2.1)	74.43 (2.2)	89.42 (2.2)	80.83 (2.2)	29.02 (0.7)	97.78 (2.7)
MMD-AE	0%	99.96 (3.0)	99.96 (2.5)	49.95 (1.3)	41.32 (0.9)	**95.87 (2.7)**	84.01 (1.8)	21.15 (0.6)	**98.56 (2.9)**
	20%	**99.99 (2.8)**	**99.99 (2.7)**	98.47 (2.9)	70.54 (1.9)	90.54 (2.7)	**84.01 (2.2)**	42.53 (0.9)	96.49 (2.1)
	30%	99.99 (2.4)	99.99 (2.8)	99.11 (2.1)	73.60 (1.7)	89.42 (2.5)	77.60 (2.2)	58.04 (1.4)	93.48 (2.4)
	50%	99.61 (2.4)	99.46 (2.4)	**99.30 (2.5)**	**78.23 (2.2)**	94.84 (2.4)	72.08 (2.1)	**67.21 (1.5)**	92.41 (2.3)
NSAE	0%	**99.99 (2.5)**	99.99 (2.0)	46.90 (1.1)	32.42 (0.8)	69.11 (1.7)	**99.90 (2.7)**	**56.05 (1.1)**	94.68 (2.0)
	20%	99.99 (3.0)	**99.99 (2.9)**	97.77 (2.8)	69.11 (1.9)	**99.90 (2.6)**	99.87 (2.1)	36.66 (0.9)	**97.71 (2.3)**
	30%	99.99 (2.6)	99.99 (2.9)	**98.70 (2.5)**	**71.29 (1.9)**	99.71 (2.7)	99.90 (2.1)	43.83 (1.2)	96.11 (2.0)
	50%	99.99 (2.4)	99.99 (2.7)	98.62 (2.1)	70.64 (2.0)	99.89 (2.3)	99.90 (2.6)	49.70 (1.1)	95.78 (2.4)
CORAL	0%	**99.99 (2.6)**	99.99 (3.0)	98.53 (2.4)	74.43 (2.2)	89.42 (2.0)	**81.13 (2.3)**	30.53 (0.8)	98.73 (2.7)
	20%	99.99 (2.2)	**99.99 (2.8)**	98.53 (2.8)	74.43 (2.2)	89.42 (2.1)	81.13 (2.1)	**42.30 (1.0)**	98.73 (2.2)
	30%	99.99 (2.4)	99.99 (2.9)	**98.86 (2.5)**	75.55 (1.9)	**99.90 (2.9)**	80.08 (2.3)	42.03 (0.9)	**98.73 (2.5)**
	50%	99.99 (2.5)	99.99 (2.7)	99.11 (2.3)	**77.21 (2.0)**	99.57 (2.7)	80.51 (2.2)	42.30 (1.0)	98.73 (2.4)
MTLS-RED	0%	99.99 (2.7)	99.99 (2.8)	98.41 (2.4)	78.92 (2.0)	78.21 (1.9)	76.04 (1.8)	73.91 (1.7)	87.67 (2.5)
	20%	99.99 (2.6)	99.99 (2.7)	98.41 (2.3)	78.92 (1.9)	78.21 (1.8)	99.75 (2.5)	68.03 (1.6)	91.83 (2.4)
	30%	99.99 (2.8)	99.99 (2.5)	98.41 (2.8)	78.21 (1.9)	53.19 (1.5)	99.87 (2.6)	73.67 (1.8)	**93.05 (2.3)**
	50%	**99.99 (2.5)**	**99.99 (2.8)**	**98.98 (2.5)**	**81.17 (2.3)**	**95.87 (2.7)**	**99.88 (2.3)**	**75.91 (1.6)**	91.20 (2.4)

ularization—enhance generalization to unseen domains. The datasets are added strategically to improve coverage of unseen domains, ensuring a broader representation. In most cases, assigning *higher weight to the MI penalty* while keeping the *reconstruction weight minimal* leads to the best generalization performance. Figure 2 demonstrates the impact of the regularization penalty on the RARE

dataset. Our results indicate that incorporating regularization—and particularly increasing the weight on the mutual information (decorrelation) penalty—leads to improved and more stable precision, recall, and AUC when training on the combined GOLDEN-EYE and SOLARIS dataset.

Table 3: We report **accuracy** (with std deviation) of the proposed and baseline methods on the Arrythmia dataset. For each case, the dataset contains equal amount of normal and anomaly samples.

Model	Percent (%)	SOURCE DOMAINS		CROSS DOMAINS		OOD DOMAINS
		VEB	BENIGN	SVEB	Q	F
MTAE	50%	99.17 (2.48)	62.94 (1.57)	60.53 (1.51)	90.41 (2.26)	60.00 (2.50)
MMD-AE	50%	98.58 (2.46)	63.63 (1.59)	44.90 (1.12)	93.33 (2.33)	89.66 (2.24)
NSAE	50%	97.18 (2.43)	72.71 (1.82)	41.70 (1.04)	80.00 (2.00)	90.17 (2.25)
CORAL	50%	97.87 (2.45)	58.24 (1.46)	69.70 (1.74)	95.26 (2.38)	73.33 (1.83)
MTLS-RED	50%	**99.34 (2.48)**	**77.49 (1.94)**	**73.48 (1.84)**	**95.52 (2.39)**	**93.33** (2.33)

In Figure 4, we visualize the latent space representations of a standard multi-task encoder-decoder model without regularization and our proposed model incorporating the MI penalty. We observe improved clustering of source, cross-domain, and target domain classes when regularization is applied, demonstrating its effectiveness in structuring the latent space. As shown in Table 2, increasing the proportion of cross-domain data during training significantly enhances classification performance across OOD datasets, such as WEB and SPOOFING attacks. Likewise, Table 3 presents the evaluation of our method on the Arrhythmia dataset, where the model is trained on the normal and VEB classes while considering all other anomaly classes—SVEB, Q as cross-domain, and F—as OOD test class. In Table 4, we evaluate our approach on a time-series dataset (EMG Gesture Recognition) and compare with the baselines.

MODEL	Domain1	Domain2	Domain3	Domain4
DIFEX	65.02 ± 2.00	66.15 ± 2.50	**64.06 ± 2.00**	62.98 ± 2.00
CORAL	52.39 ± 2.00	52.51 ± 2.50	53.89 ± 2.00	57.06 ± 2.00
MTL-RED	**66.41 ± 3.40**	**66.30 ± 2.50**	55.92 ± 2.00	**65.82 ± 2.50**

Table 4: Performance (accuracy %) of MTL-RED, DIFEX, CORAL with EMG time series dataset divided into 4 domains each consisting of 6 classes for all the 9 persons.

5 Conclusion

Our paper addresses the challenge of detecting novel and out-of-distribution (OOD) anomalies through domain generalization techniques. By training on mul-

tiple source and cross-domain datasets with distinct correlation structures, we aim to increase the coverage to generalize to unseen anomaly classes. Subsequently, guided by the principle of relevant information preservation (PRI), our regularization steers the cross-entropy loss in latent space to retain essential features to achieve domain generalization. Real-world cybersecurity and healthcare datasets often exhibit different correlation patterns (varying joint distributions) among the different classes, which can be exploited to increase the coverage for extrapolation to new, unseen domains. Future work will further explore methods for latent space anomaly detection.

Acknowledgment

We gratefully acknowledge the support of the Virginia Tech National Security Institute (VTNSI) and the Deloitte & Touche LLP, USA, for supporting this research. We also extend our sincere thanks to our collaborators at Deloitte — Ajay Kumar, Alison Hu, Sanmitra Bhattacharya, and Edward Bowen for their insightful contributions.

References

1. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2551–2559 (2015)
2. Sathya, R., Sekar, K., Ananthi, S., Dheepa, T.: Adversarially Trained Variational Auto-Encoders With Maximum Mean Discrepancy based Regularization. In: 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), pp. 1–6. IEEE (2022)
3. Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T.G.J., LeCun, Y.: An information-theoretic perspective on variance-invariance-covariance regularization. arXiv preprint arXiv:2303.00633 (2023)
4. Jeon, E., Ko, W., Yoon, J.S., Suk, H.I.: Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI. IEEE Transactions on Neural Networks and Learning Systems **34**(2), 739–749 (2021)
5. Venkataramanan, S., Psomas, B., Kijak, E., Amsaleg, L., Karantzalos, K., Avrithis, Y.: It takes two to tango: Mixup for deep metric learning. arXiv preprint arXiv:2106.04990 (2021)
6. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
7. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3964–3973 (2018)
8. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. Advances in Neural Information Processing Systems **35**, 32598–32611 (2022)
9. Wang, Wei, Vincent W. Zheng, Han Yu, and Chunyan Miao. "A survey of zero-shot learning: Settings, methods, and applications." ACM Transactions on Intelligent Systems and Technology (TIST) 10, no. 2 (2019): 1-37.

10. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning, pp. 20827–20840. PMLR (2022)
11. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
12. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2018)
13. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2507–2516 (2019)
14. Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: ECCV 2020, pp. 301–318. Springer International Publishing (2020)
15. Liang, H., Zhang, Q., Dai, P., Lu, J.: Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9424–9434 (2021)
16. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D.: Matching networks for one shot learning. Advances in Neural Information Processing Systems **29** (2016)
17. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)
18. Lu, W., Wang, J., Li, H., Chen, Y., Xie, X.: Domain-invariant feature exploration for domain generalization. arXiv preprint arXiv:2207.12020 (2022)
19. Vuorio, R., Sun, S.H., Hu, H., Lim, J.J.: Multimodal model-agnostic meta-learning via task-aware modulation. Advances in Neural Information Processing Systems **32** (2019)
20. Canadian Institute for Cybersecurity: Public datasets for intrusion detection and anomaly detection, including CSE-CIC-IDS2018, IoT Dataset, IoMT Dataset, and Arrhythmia Dataset. Available at https://www.unb.ca/cic/datasets/, last accessed 2025/02/14.
21. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
22. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV 2020, pp. 124–141. Springer International Publishing (2020)
23. Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain generalization: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
24. Li, Haoliang, 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. 2018.
25. Sung, Flood, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M. Hospedales. "Learning to compare: Relation network for few-shot learning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199-1208. 2018.
26. Yu, Shujian, Francesco Alesiani, Xi Yu, Robert Jenssen, and Jose Principe. "Measuring dependence with matrix-based entropy functional." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10781-10789. 2021.

27. Wang, Jindong, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. "Generalizing to unseen domains: A survey on domain generalization." IEEE Transactions on Knowledge and Data Engineering (2022).
28. Li, Ya, 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. 2018.
29. Jin, Xin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. "Feature alignment and restoration for domain generalization and adaptation." arXiv preprint arXiv:2006.12009 (2020).
30. Erfani, Sarah, Mahsa Baktashmotlagh, Masud Moshtaghi, Xuan Nguyen, Christopher Leckie, James Bailey, and Rao Kotagiri. "Robust domain generalisation by enforcing distribution invariance." In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp. 1455-1461. AAAI Press, 2016.
31. Rusu, Andrei A., Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. "Meta-learning with latent embedding optimization." arXiv preprint arXiv:1807.05960 (2018).
32. Li, Da, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. "Learning to generalize: Meta-learning for domain generalization." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.
33. Seo, Seonguk, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. "Learning to optimize domain specific normalization for domain generalization." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pp. 68-83. Springer International Publishing, 2020.
34. Sun, Baochen, and Kate Saenko. "Deep coral: Correlation alignment for deep domain adaptation." In Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, pp. 443-450. Springer International Publishing, 2016.
35. Tishby, N., Pereira, F. C., Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.
36. Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. arXiv preprint arXiv:2211.11719, 2022.

6 Joint Optimization and the Principle of Relevant Information Preservation (PRI)

Our approach is motivated by Tishby's Principle of Relevant Information (PRI), which states that an optimal representation should preserve only the information in the input that is necessary for the task at hand, while discarding irrelevant details. In our context, the goal is to learn a latent representation Z from the input X that is both predictive of the target Y and minimally influenced by spurious correlations present in X. To achieve this, we jointly optimize a loss function that combines three key components:

1. **Cross-Entropy Loss** ($\mathcal{L}_{\text{CE}}$): This term ensures that the latent representation Z is discriminative enough to accurately predict the target Y.

2. **Reconstruction Loss** ($\mathcal{L}_{\text{recon}}$): This term (e.g., mean squared error) forces Z to retain sufficient information to reconstruct the input X, thereby preventing excessive compression.
3. **Mutual Information Penalty** ($\mathcal{L}_{\text{MI}}$): By penalizing the mutual information between X and Z, this term encourages the model to discard spurious and domain-specific correlations, leading to a more invariant and disentangled latent space.

The overall objective can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \lambda_{\text{MI}}\,\mathcal{L}_{\text{MI}}, \tag{7}$$

where λ_{recon} and λ_{MI} are hyperparameters that balance the trade-off between reconstruction fidelity and the strength of the decorrelation (compression) penalty.

In the framework of PRI (proposed by Tishby and later on implemented in various contexts), we aim to minimize the mutual information between X and Z while maintaining high mutual information between Z and Y. This idea is often expressed as:

$$\mathcal{L}_{\text{PRI}} = I(X;Z) - \beta\, I(Y;Z), \tag{8}$$

where:

- $I(X;Z)$ quantifies the total information that the latent representation Z retains about X.
- $I(Y;Z)$ measures the information in Z that is useful for predicting Y.
- β is a parameter controlling the trade-off between compression (minimizing $I(X;Z)$) and predictive power (maximizing $I(Y;Z)$).

In practice, our joint loss in Equation (7) serves as a proxy for the PRI objective in Equation (8):

- The $\mathcal{L}_{\text{CE}}$ term drives Z to retain information relevant to Y (i.e., maximizing $I(Y;Z)$).
- The combination of $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{MI}}$ encourages Z to compress X by preserving only the necessary information and discarding spurious correlations, effectively minimizing $I(X;Z)$.

Moreover, our implementation of the mutual information penalty is based on the Renyi entropy of kernel matrices computed from X and Z, with kernel bandwidths s_x and s_y that are adjusted during training. This enables the model to learn an optimal level of decorrelation, ensuring that the latent space does not overfit to domain-specific artifacts while still preserving the relevant structure needed for accurate classification and reconstruction. In summary, our joint optimization framework, which integrates classification, reconstruction, and decorrelation, is a practical instantiation of the PRI principle. By carefully balancing these objectives, our model is guided to learn a compressed yet task-relevant latent representation that generalizes effectively across domains.