Learning by Erasing: Conditional Entropy Based Transferable Out-of-Distribution Detection

Meng Xing[1,3], Zhiyong Feng[1], Yong Su[2], Changjae Oh[3]
[1]College of Intelligence and Computing, Tianjin University
[2]Tianjin Normal University
[3]Centre for Intelligent Sensing, Queen Mary University of London
{xingmeng, zyfeng, suyong}@tju.edu.cn, c.oh@qmul.ac.uk

Abstract

Detecting Out-of-distribution (OOD) inputs is crucial to deploying machine learning models to the real world safely. However, existing OOD detection methods require an in-distribution (ID) dataset to retrain the models. In this paper, we propose a Deep Generative Models (DGMs) based transferable OOD detection that does not require retraining on the new ID dataset. We first establish and substantiate two hypotheses on DGMs: DGMs are prone to learn low-level features rather than high-level semantic information; the lower bound of DGM's log-likelihoods is tied to the conditional entropy between the model input and target output. Drawing on the aforementioned hypotheses, we present an innovative image-erasing strategy, which is designed to create distinct conditional entropy distributions for each ID dataset. By training a DGM on a complex dataset with the proposed image-erasing strategy, the DGM could capture the discrepancy of conditional entropy distribution for varying ID datasets, without re-training. We validate the proposed method on the five datasets and show that, without retraining, our method achieves comparable performance to the state-of-the-art group-based OOD detection methods. The project codes will be open-sourced on our project website.

Introduction

Deep neural networks (DNNs) have demonstrated their potential in solving various safety-related computer vision tasks (Wang, Shi, and Yeung 2016), such as autonomous driving (Casas, Sadat, and Urtasun 2021) and healthcare (Kim et al. 2021). However, DNNs tend to yield confident but incorrect predictions for the distribution-mismatched examples (Nguyen, Yosinski, and Clune 2015; Sensoy, Kaplan, and Kandemir 2018; Shekhovtsov and Flach 2019), and results in serious consequences, e.g., accidents by autonomous vehicles (Times 2018) and incorrect diagnosis in healthcare (BBC 2020). Therefore, determining whether inputs are out-of-distribution (OOD) is an important task to safely deploy machine learning models to the real world.

OOD detection can be performed using labeled data by utilizing output characteristics (Hsu et al. 2020), training dynamics (Huang et al. 2021), adversarial training (Lakshminarayanan, Pritzel, and Blundell 2017; Bevandic et al. 2018; Huang and Li 2021), and metric learning (Lee et al. 2018b; Zaeemzadeh et al. 2021; Ming et al. 2023). Since it is time-consuming and laborious to obtain labeled data in real scenarios, as an alternative, Deep Generative Models (DGMs) have been used to capture the sample distribution of In-Distribution (ID) datasets (Serrà et al. 2020). However, most DGMs-based methods focus on elaborating architectures (Ren et al. 2019; Serrà et al. 2020), designing loss functions (Xiao, Yan, and Amit 2020) or statistical models (Zhang et al. 2020; Jiang, Sun, and Yu 2022), targeting the specific feature representation or data distribution of ID samples (Sun et al. 2023), i.e., need retraining to adapt to the normal pattern of the new ID datasets. This motivates the following unexplored question: *How can we make OOD detection transferable across new ID datasets?*

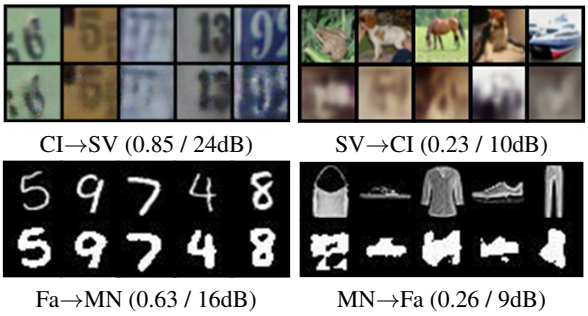

Figure 1: Original images (top row) and their reconstruction results (bottom row) by a pre-trained DGM. The model pre-trained on CI (CIFAR10) / Fa (FashionMNIST) can reconstruct SV (SVHN) / MN (MNIST) samples well, but not *vice versa*. Reconstruction performances on the test set of the target datasets are evaluated with (SSIM↑ / PSNR↑).

In this paper, we aim to achieve transferable OOD detection based on the following two key hypotheses: *1) The DGMs are prone to learn low-level features, rather than semantic information (Kirichenko, Izmailov, and Wilson 2020).* Following the experimental setup in (Xiao, Yan, and Amit 2020), we use the Variational Auto-Encoder to reconstruct the input image and show some results comparisons in Figure 1. Figure 1 demonstrates that a DGM pre-trained on a complex dataset, which includes diverse semantic categories and a complex image texture, can cap-

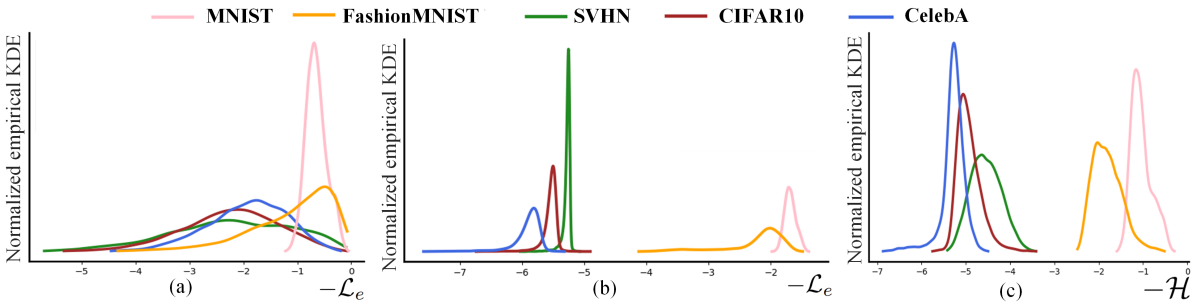

Figure 2: The (a) and (b) are the DGM's negative log-likelihood distribution on different datasets. The model of (a) is trained by reconstructing the input, while the model of (b) is trained by generating the erased patch based on its surrounding information. The real negative conditional entropy distribution between the erased patch and its surrounding is given in (c). The DGM is an auto-encoder proposed in this paper and is trained with ImageNet. The image size is 32×32, and the erased image patch in (b) and (c) is at the center of the image with the size of 16×16. Kernel Density Estimation (KDE) is used to estimate the probability distribution.

ture the distribution of simple datasets, but not *vice versa*. This means that the DGMs pre-trained on a complex dataset can approach the lower bounds of negative-log-likelihoods of simple datasets without retraining.

2) In the DGMs, the lower bound of negative-log-likelihoods is determined by the conditional entropy between the model input and target output. We give supporting experimental results in Figure 2 and theoretically demonstrate this hypothesis in *Motivation*. The log-likelihood distributions of all datasets in Figure 2(a) are approaching 0 since the conditional entropy between the model input and output is 0 in this experiment setting. This result explains why traditional DGMs cannot be used directly for OOD detection(Serrà et al. 2020). In contrast, the log-likelihood distributions of five datasets in Figure 2(b) are significantly different from each other and the distribution discrepancy is consistent with real conditional entropy distribution in Figure 2(c). Therefore, we can assign an exclusive conditional entropy distribution for each dataset by designing an appropriate image-erasing strategy, which is an indispensable prerequisite for achieving transferable OOD detection.

Motivated by the proven hypotheses, we propose a novel Conditional Entropy based Transferable OOD detection (CETOOD). Specifically, we first propose an image-erasing strategy that creates exclusive conditional entropy distribution for different datasets by considering the erased patch and its surrounding information as the content and condition. Subsequently, we design the Uncertainty Estimation Network (UEN), which estimates the Maximum A Posteriori of generating the erased patch by reconstructing the surrounding information and generating the erased patch. Finally, we train the UEN on ImageNet (Deng et al. 2009) dataset, affording our model approaching the lower bounds of negative-log-likelihoods on different ID datasets, which reflects their distribution discrepancy of conditional entropy. In the experiment, we demonstrate that our method achieves comparable performance with state-of-the-art methods in group-based OOD detection. More importantly, our pipeline drastically curtails the time and memory cost of model deployment due to its transferability and concise network architecture. In summary, our contributions are as follows:

- We introduce the concept of conditional entropy into OOD detection for model transferability, and theoretically demonstrate the lower bound of negative-log-likelihoods in DGMs is determined by the conditional entropy between the model input and target output.
- We propose a transferable OOD detection method (CETOOD), which captures the distribution discrepancy of conditional entropy of different ID datasets to achieve transferable OOD detection.
- We demonstrate the effectiveness and lightweight of the proposed method through extensive comparisons with state-of-the-art techniques, across different datasets.

Related Work

Some **Classifier-based approaches** detect OOD samples by utilizing the statistical characteristic of class probabilities. Hendrycks *et al.* (Hendrycks and Gimpel 2017) propose maximum softmax probability as a baseline for OOD detection in deep neural network (DNN) algorithms, and ODIN (Liang, Li, and Srikant 2018) further enhance the performance by using temperature scaling and adding small perturbations on ID inputs. Since the distribution of OOD data is not available, some methods have explored using synthesized data from generative adversarial networks (GANs) (Lee et al. 2018a) or using unlabeled data (Hendrycks, Mazeika, and Dietterich 2019; Mohseni et al. 2020) as auxiliary OOD training data, which allows the model to be explicitly regularized by fine-tuning, producing lower confidence on anomalous examples. In addition to these softmax-classification-based frameworks, recently, researchers focus on the feature embedding of the model. With the observation that the unit activation patterns of a particular layer show a significant difference between ID and OOD data, Djurisic *et al.* (Djurisic et al. 2023) utilize feature transformation to generate the OOD score. Similarly, some methods exploit hyperspherical embeddings (Ming et al. 2023) or

cosine similarity (Nguyen et al. 2023) between features to promote strong ID-OOD separability. Despite the promising results, classification-based approaches show limitations on the non-labeled tasks.

As an alternative, most **DGM-based OOD detection** methods separate the ID and OOD samples by exploiting the inductive bias of DGMs, including background statistics (Ren et al. 2019; Cai and Li 2023), inputs complexity (Serrà et al. 2020) and low-level features (Sun et al. 2023). Xiao *et al.* (Xiao, Yan, and Amit 2020) propose the Likelihood Regret, which is a log-ratio between the likelihood of input obtained by posteriori distribution and approximated by VAE, to detect OOD samples. Serrà *et al.* (Serrà et al. 2020) design a complexity estimate score and utilize the subtraction between negative log-likelihoods and the complexity estimate score to detect OOD inputs. Kirichenko *et al.* (Kirichenko, Izmailov, and Wilson 2020) prove through experiments that what DGMs learn from images is local pixel correlation and local geometric structure rather than semantic information. Therefore, Sun *et al.* (Sun et al. 2023) utilizes sample repairing to encourage the generative model to focus on semantics instead of low-level features. A recent work (Zhang, Goldstein, and Ranganath 2021) has shown that for the point-based OOD detection method, a perfect model can perform worse than a falsely estimated one when the ID and OOD data are overlapped.

Therefore, **Group-based OOD detection methods** utilize the distribution characteristics of grouped inputs for OOD detection. Most group-based methods consider either the raw input or a certain representation of samples for OOD detection. Nalisnick *et al.* (Nalisnick et al. 2019) propose an explicit test for typicality employing a Monte Carlo estimate of the empirical entropy. As an alternative, exploit data representations in the latent space can also be utilized to achieve OOD. Zhang *et al.* (Zhang et al. 2020) find that the representations of inputs in DGMs can be approximated by fitted Gaussian and the distance between the distribution of representations of inputs and prior of the ID dataset can be utilized to detect OOD samples. Jiang *et al.* (Jiang, Sun, and Yu 2022) propose to compare the training and test samples in the latent space of a flow model. However, these methods require retraining when encountering new ID datasets, which is computationally expensive and time-consuming.

Motivation

In this section, we demonstrate the relationship between the lower bound of DGM's negative-log-likelihoods and the conditional entropy between model input and target output. We take the grayscale image as an example, which can be extended to RGB easily.

Given an image pair (A, B), we can calculate the uncertainty of random variable B given random variable A, i.e., the conditional entropy $H(B|A)$ as follows:

$$\begin{aligned} H(B|A) &= -\Sigma_{i=0}^{N_B}\Sigma_{j=0}^{N_A} P(A_j, B_i) log(P(B_i|A_j)) \\ &= -\Sigma_{i=0}^{N_B} P(B_i) log(P(B_i|A)) \end{aligned}$$

where A_j and B_i are the pixel value at locations j and i of images A and B. N_A and N_B are the number of pixel of images A and B.

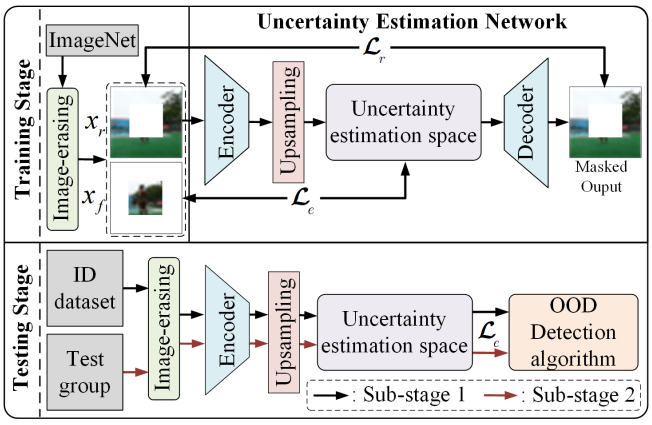

Figure 3: The pipeline of the proposed CETOOD.

For image generation, given the model input A, target output B and a pre-trained DGM (parameters: Z), the Maximum A Posteriori (MAP) of generating output can be estimated as follows:

$$MAP = \arg\min_{Z} KL(P(B|Z)P(Z) \;||\; P(B|A)P(A))$$

According to the information bottleneck theory (Tishby, Pereira, and Bialek 2000), the lower bound of the negative-log-likelihoods of DGM can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{lower\ bound} &= -(log(P(B|A)) + log(P(A))) \\ &= -(\Sigma_{i=0}^{N_B} log(P(B_i|A))/N_B) - log(P(A)) \\ &= \underbrace{-(\Sigma_{i=0}^{N_B} P(B_i) log(P(B_i|A)))}_{H(B|A)} - log(P(A)) \end{aligned}$$

where $P(B_i|A)$ is modeled by the pre-trained DGM.

Therefore, given A as input, the conditional entropy between A and B would determine the lower bound of DGM's negative-log-likelihoods.

Method

The proposed framework consists of the image-erasing strategy, UEN, and OOD detection algorithm, as shown in Figure 3.

Image-Erasing Strategy

To create exclusive conditional entropy distribution for different datasets, we design an image-erasing strategy that divides the image into the erased patch and its surrounding information. Conditional entropy is a measure of the information difference between the image's erased patch and its surrounding. Due to semantic differences existing between different datasets, exclusive conditional entropy distributions can be generated for different datasets by erasing the most semantically meaningful regions. We empirically choose to erase the center of the input, but also propose other erasing strategies for comparison. The details of the image-erasing strategies are shown in Figure 4.

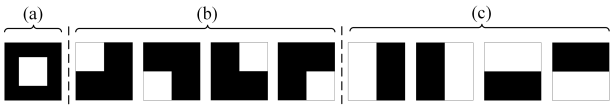

Figure 4: The center (a), corner (b) and side (c) of the image is erased, which is indicated with white color.

With the original image x, the model input (surrounding information) x_r and output (erased patch) x_f are generated as follows:

$$x_r = Mask(x), \text{ with } x_f = x - x_r, \tag{1}$$

where $Mask(x)$ indicates putting a mask on x.

Uncertainty Estimation Network

To capture the distribution discrepancy of conditional entropy for different datasets, we propose UEN, a concise auto-encoder, as shown in Figure 3. To estimate the MAP of generating the erased patch, UEN needs to calculate the probability of generating the target output directly based on the model input. We assume that the pixel values at each position of the image conform to a continuous distribution, and all parameters of the distribution depend on the model input. Inspired by PixelCNN++ (Salimans et al. 2017), we use the mixed logistic distribution as the above continuous distribution and name the feature space on which the parameters of the distribution depend as uncertainty estimation space, Z:

$$Z \sim \sum_{k=1}^{K} \pi_k \frac{e^{-(x-\mu_k)/\gamma_k}}{\gamma_k(1+e^{-(x-\mu_k)/\gamma_k})^2}, \tag{2}$$

where K is the number of components in the mixed logistic distribution, π_k is the weight of each component, μ_k and γ_k are the shape and position parameters of the logistic distribution, respectively (π_k, μ_k and γ_k are learnable parameters, where $k = \{1, 2, \ldots, K\}$).

Given the erased patch, x_f, the likelihood of each discretized pixel value can be directly calculated as follows:

$$P(x_f|Z) = \sum_{k=1}^{K} \pi_k [\sigma(\frac{x_f + \frac{1}{255} - \mu_k}{\gamma_k}) - \sigma(\frac{x_f - \frac{1}{255} - \mu_k}{\gamma_k})], \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, and we set K to 10 in this paper. For RGB images, we only allow linear dependence between three-channel pixel values.

The encoder consists of three parallel multi-layer convolutional branches with different kernel sizes. Upsampling layers are designed to ensure the size of the feature map in the uncertainty estimation space is consistent with the input. The deep feature in uncertainty estimation space is further mapped into the image domain by a decoder.

To ensure that no surrounding information is lost in the process of constructing the uncertainty estimation space, i.e, $P(x_f|Z) \approx P(x_f|x_r)$, we design the reconstruction loss, $\mathcal{L}_r$, as follows:

$$\mathcal{L}_r = \|x_r - o_r\|^2, \tag{4}$$

Algorithm 1: OOD Detection Algorithm

Require: Z: pre-constructed uncertainty estimation space; $X^* = \{x_1^*, x_2^*, \ldots x_N^*\}$: all of ID samples; $X = \{x_1, x_2, \ldots x_m\}$: a batch of of test samples; $Mask()$: the function of erasing image patch; t: threshold.

$i \leftarrow 1$
while $i \leq N$ **do**
  $x_{if}^* = x_i^* - Mask(x_i^*)$
  $L^*[i] = \mathcal{L}_e(x_{if}^*|Z); i \leftarrow i + 1$
end while
$P(L^*) = KDE(L^*)$
while $testset \neq \varnothing$ **do**
  $j \leftarrow 1$
  while $j \leq m$ **do**
    $x_{jf} = x_j - Mask(x_j)$
    $L[j] = \mathcal{L}_e(x_{jf}|Z); j \leftarrow j + 1$
  end while
  $P(L) = KDE(L)$
  $k = KL(P(L)\|P(L^*))$
  if $k > t$ **then**
    return X is out-of-distribution data.
  else
    return X is in-distribution data.
  end if
  reload X
end while

where $o_r = Mask(o)$ is the masked output and o is the model output.

To highlight the distribution discrepancy, the generation loss, $\mathcal{L}_e$, which measures the posterior probability of generating the erased patch, is presented as:

$$\begin{aligned} \mathcal{L}_e &= -log_2[P(x_f|Z)] \\ &= -\sum_{i=1}^{N} log_2[P(x_{fi}|Z)]/N_f, \end{aligned} \tag{5}$$

where i is the pixel location, x_{fi} is the pixel value at location i and N_f is the number of pixels in the erased patch, x_f.

$\mathcal{L}_e$ encourages UEN to narrow the log-likelihood distribution gap between the samples that contain similar semantic discrepancies. The final loss function is as follows:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_r + (1 - \lambda)\mathcal{L}_e, \tag{6}$$

where λ is used to balance the effect between $\mathcal{L}_r$ and $\mathcal{L}_e$.

OOD Detection Algorithm

Algorithm 1 shows the proposed OOD detection using the pre-trained uncertainty estimation network. Given all ID samples $X^* = \{x_1^*, x_2^*, \ldots x_N^*\}$ and image-erasing strategy $Mask()$, we first utilize Kernel Density Estimation (KDE) to obtain the distribution of log-likelihood for ID dataset. Then, in the same way, given a set of test samples $X = \{x_1, x_2, \ldots, x_n\}$, we estimate the distribution of log-likelihood on the test group. Finally, we measure the estimated total correlation between the test group and the ID samples by using KL-divergence, and determine the test group as the OOD data if there exists a significant distribution discrepancy.

Methods (*Retraining*)			DOCR-TC-M (*Required*)		Ty-test (*Required*)		RF-GM (*Required*)		Ours ***(Not required)***	
GS	ID	OOD	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
5	MNIST	F-MNIST	-	-	-	-	-	-	99.2	97.2
	F-MNIST	MNIST	**100.0**	**100.0**	95.5	92.1	99.0	99.0	93.8	89.5
	SVHN	CIFAR10	99.7	99.7	**100.0**	**100.0**	89.0	93.0	99.2	97.5
		CelebA	**100.0**	**100.0**	100.0	100.0	92.0	94.0	98.9	96.0
	CelebA	CIFAR10	91.6	91.9	5.7	31.2	**92.0**	**93.0**	91.2	86.8
		SVHN	**100.0**	**100.0**	83.1	80.1	97.0	96.0	91.2	87.2
	CIFAR10	SVHN	99.0	**99.6**	98.6	99.3	88.0	83.0	**99.7**	85.6
		CelebA	**100.0**	**100.0**	100.0	100.0	76.0	77.0	98.6	92.6
10	MNIST	F-MNIST	-	-	-	-	-	-	100.0	100.0
	F-MNIST	MNIST	**100.0**	**100.0**	99.4	99.3	99.0	99.0	99.1	96.3
	SVHN	CIFAR10	100.0	100.0	100.0	100.0	95.0	98.9	**100.0**	**100.0**
		CelebA	100.0	100.0	100.0	100.0	98.0	99.0	**100.0**	**100.0**
	CelebA	CIFAR10	99.2	**99.3**	0.9	30.7	98.0	99.0	**99.4**	93.9
		SVHN	**100.0**	**100.0**	91.6	90.5	100.0	100.0	99.0	94.6
	CIFAR10	SVHN	**100.0**	**100.0**	99.9	100.0	99.0	98.0	99.3	99.7
		CelebA	**100.0**	**100.0**	100.0	100.0	89.0	90.0	99.9	99.9

Table 1: The OOD detection results with different group size (GS) on five different datasets. Unlike other methods, our transferable method does not require retraining on the ID dataset.

Experiments

Implementation Details

All three parallel encoder branches consist of multiple convolution and upsampling layers with different kernel sizes (3×3, 5×5 and 7×7). A shared convolutional layer with a kernel size of 1×1 is utilized to transform the features from 3 parallel encoder branches into uncertainty estimation space. The decoder consists of two convolutional layers with kernel size of 3×3. We set the batch size and learning rate to 64 and 10^{-5}, respectively. λ is empirically set to 0.8. We trained the network for 250 epochs, taking about 48.29 hours. We conduct all experiments on a single NVIDIA GPU 3080 that follows the experimental setup of the baseline methods.

Experimental Setting

Datasets We train our model on ImageNet32 (Deng et al. 2009) and validate our model on different ID datasets, including MNIST (LeCun et al. 1998), FashionMNIST (Xiao, Rasul, and Vollgraf 2017), SVHN (Netzer et al. 2011), CelebA (Liu et al. 2015) and CIFAR10 (Krizhevsky and Hinton 2009). All the inputs are resized to 32×32 to fit the input size of UEN. We transform the grayscale image into an RGB image by replicating the channel.

Metrics We use threshold-independent metrics: the area under the receiver operating characteristic curve (AUROC) (Davis and Goadrich 2006) and the area under the precision-recall curve (AUPR) to evaluate our method. We consider OOD data and ID data as positive and negative ones for detection, respectively. Unless noted otherwise, we calculate the False Positive Rate (FPR) of the detector when the threshold is set at 95% TPR. We randomly select 10k samples from the test set of the target dataset. We generate test sample groups according to group size gs. For the fair comparison, we generate the test set 2 times and test groups 5 times then report the averaged result.

OOD Detection

To evaluate the robustness of our method, we utilize five different datasets as ID datasets and test each of them on one (MNIST or FashionMNIST) or three (SVHN, CelebA and CIFAR10) different disjoint OOD datasets. The obtained performance for OOD detection and comparison with three baselines including the Ty-test (Nalisnick et al. 2019), DOCR-TC-M (Zhang et al. 2020) and RF-GM (Jiang, Sun, and Yu 2022) are shown in Table 1. We utilize the three methods as our baselines as they outperform other existing group-based OOD detection methods. As shown in Table 1, our method can achieve competitive performance compared with the SOTA methods. Our method achieves higher AUROC compared to RF-GM across various detection scenarios, especially, our method outperforms RF-GM 22.6% AUROC when detecting CelebA from CIFAR10 with 5 as group size. Likewise, our method shows 0.7% higher AUROC compared to DOCR-TCM when detecting SVHN from CIFAR10 with 5 as group size. Notably, compared to the baseline methods, our framework does not require retraining when deployed on new ID datasets.

Deployment Cost Analysis

In order to comprehensively analyze the performance of our model, we compare the time (training time) and space complexity (memory cost of network parameters) of our approach with that of the baseline methods. Due to both DOCR-TC-M and RF-GM are based on the flow model, we choose the DOCR-TC-M with better performance as a baseline. The experiment settings for DOCR-TC-M (Zhang et al. 2020) and Ty-test (Nalisnick et al. 2019) are consistent with the original papers. The training time and memory cost comparison are shown in Figure 5. Our method does not require retraining and only needs to calculate the DGM's likelihood distribution of the new ID dataset in the testing stage. Therefore, the time cost of model deployment can be greatly re-

ID	OOD	(a) Group size			(b) Erasing strategy			(c) Erasing strategy ($\mathcal{H}$)		
		20	50	100	corner	side	center	corner($\mathcal{H}$)	side($\mathcal{H}$)	center($\mathcal{H}$)
MNIST	F-MNIST	100.0	100.0	100.0	95.2	99.7	100.0	99.7	99.9	99.9
F-MNIST	MNIST	99.9	100.0	100.0	91.4	97.2	99.1	94.2	97.7	99.6
SVHN	CelebA	100.0	100.0	100.0	100.0	100.0	100.0	99.7	99.7	99.9
	CIFAR10	100.0	100.0	100.0	99.9	100.0	100.0	99.5	99.1	99.9
CelebA	SVHN	99.9	100.0	100.0	89.4	90.1	99.0	100.0	100.0	100.0
	CIFAR10	99.8	100.0	100.0	64.8	63.7	99.4	56.3	54.3	86.3
CIFAR10	SVHN	99.6	99.9	100.0	90.3	95.2	99.3	100.0	100.0	100.0
	CelebA	100.0	100.0	100.0	59.3	58.7	99.9	51.2	51.0	89.6
ID	OOD	(d) Loss function			(e) Training set					
		$\mathcal{L}_e$	$\mathcal{L}_r$	$\mathcal{L}_{total}$	MNIST	F-MNIST	SVHN	CelebA	CIFAR10	ImageNet
MNIST	F-MNIST	99.7	99.9	100.0	100.0	99.2	99.9	100.0	99.8	100.0
F-MNIST	MNIST	93.3	98.4	99.1	99.7	97.9	97.7	98.5	98.1	99.1
SVHN	CelebA	61.4	77.6	100.0	61.6	74.6	100.0	49.1	100.0	100.0
	CIFAR10	62.0	76.1	100.0	49.2	90.2	99.9	99.9	78.0	100.0
CelebA	SVHN	64.1	82.0	99.0	88.6	71.8	97.3	58.5	98.4	99.0
	CIFAR10	70.9	79.5	99.4	74.7	65.2	77.5	90.5	97.8	99.4
CIFAR10	SVHN	62.8	78.1	99.3	75.0	88.9	96.9	99.0	68.4	99.3
	CelebA	68.1	77.5	99.9	46.4	65.1	63.2	68.4	99.6	99.9

Table 2: Model performance with different hyperparameters and training variations (group size for b-e is 10).

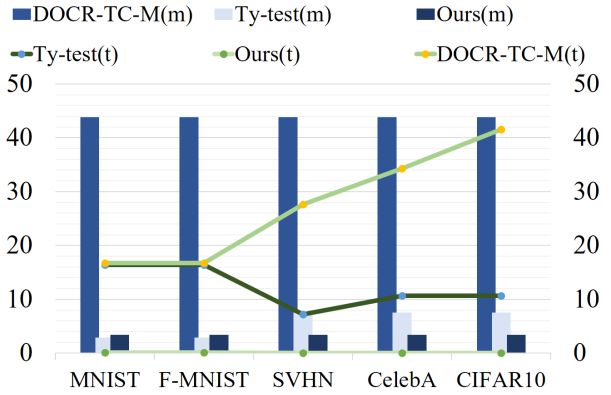

Figure 5: The time (t, hours) and space (m, FLOPs) complexity comparison between our method and the baseline approaches.

duced. Our model needs 48.3 hours to be pre-trained on ImageNet, which is still less than the time cost of training the baseline methods on all ID datasets. In addition, the space complexity comparison in Figure 5 shows that the memory cost of our model is significantly lower than the baseline methods.

Ablation Study

Effect of group size Table 2(a) reports the model performance with different group sizes. Experiment results demonstrate that the group size only has a slight impact on model performance and it is sufficient to ensure the performance of the model with the group size higher than 5.

Effect of image-erasing strategy To analyze the effect of the image-erasing strategy, we use three image-erasing strategies to train the model. Note that the same image-erasing strategy is applied to both training the model and OOD detection. For the image-erasing strategy with different variations, we calculate the average of all the variations. We tabulate the model performance in Table 2(b), denoted as corner, side and center. The experimental results indicate that the center strategy can significantly improve the detection performance in some scenarios (CelebA versus CIFAR10 and CIFAR10 versus CelebA), but only slight performance improvement is observed in other scenarios. To explore the reasons for poor robustness in performance improvement, we feed the real conditional entropy under different image-erasing strategies into Algorithm 1 and calculate the OOD detection results, as shown in Table 2(c). The experimental results show that in the scenarios with slight performance improvement, the corner and side strategies can create exclusive conditional entropy distributions for different datasets. The above experimental results support hypothesis 2), i.e., the detection performance of the model depends on the conditional entropy distribution discrepancy between different datasets.

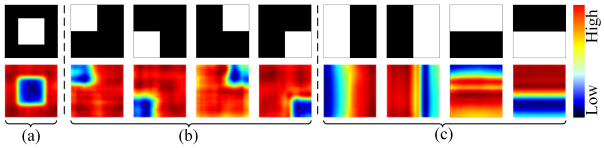

Figure 6: Top: Different image-erasing strategies. Bottom: The averaged likelihood heatmaps of all CIFAR10 test samples generated by our model with image-erasing strategies.

In addition, as the $\mathcal{L}_e$ encourages UEN to narrow the log-likelihood distribution gap between the samples with similar semantic discrepancies, the detection performance of the model should be superior to the detection results based on real conditional entropy. However, in some experimental settings, the experimental results do not match expectations. For example, when detecting SVHN from CIFAR10, the performance of the model decreases compared to the

detection results based on real conditional entropy, and the performance degradation caused by different image-erasing strategies varies significantly. To investigate the impact of the image-erasing strategy on the conditional entropy capturing, we presented the averaged likelihood heatmaps of all samples in the CIFAR10 dataset under different image-erasing strategies, as shown in Figure 6. The expected experimental results should be like Figure 6(a), the blue region in the bottom heatmap is aligned with the white region in the top sketch map, which indicates the information discrepancy between the erased patch and its surrounding can be captured by the proposed model effectively. However, in the corner (Figure 6(b)) and side (Figure 6(c)) strategies, the blue regions in the heatmaps are much smaller than their corresponding white regions. The results demonstrate that the model with these two erasing strategies could generate partial information about the target output. In other words, the model's ability to capture conditional entropy is affected by the value of conditional entropy, the larger the better.

Effect of loss functions We show the quantitative comparison results of different model objectives in Table 2(d). We also show the feature visualization of samples when the model is trained with different model objectives in Figure 7. Experimental results in Table 2(d) show that the performance of $\mathcal{L}_r$ consistently outperforms $\mathcal{L}_e$ across different datasets, which demonstrates ensure $P(x_f|Z) \approx P(x_f|x_r)$ plays a major role in capturing the inter-dataset distribution discrepancy (the conclusion is consistent with the results in Figure 7 (b) and (c)). In addition, the comparison between Figure 7 (c) and (d) shows that $\mathcal{L}_e$ reduces the distribution variance of each dataset, thus further increasing the distribution discrepancy. The results demonstrate that $\mathcal{L}_r$ ensures the model captures the condition (surrounding information) of the conditional entropy, while $\mathcal{L}_e$ encourages the model to capture the content (erased patches), and the complementarity between them helps to accurately capture the conditional entropy.

Effect of training set To analyze the effect of the training set, we train our model using training sets of 6 datasets, including MNIST, F-MNIST, SVHN, CelebA, CIFAR10 and ImageNet. As shown in Table 2(e), the results show that the model performance improves with the increases in training data's complexity and achieves optimal performance on ImageNet. The experimental results support hypothesis 1), i.e., the DGMs learn low-level features rather than semantic information. As the same image-erasing strategy is used among 6 experiment settings, the experimental results also demonstrate that conditional entropy capturing is affected by the complexity of training data. Highly complex training data helps the model better capture the conditional entropy of generating the erased patch from its surrounding.

Limitation To further explore the potential of CETOOD, we utilize the model that is pre-trained on ImageNet to distinguish CIFAR100 (Krizhevsky and Hinton 2009) and CIFAR10. We also feed the real conditional entropy into Algorithm 1 for OOD detection. The center image-erasing strategy is used in both experiments. The reason for poor model

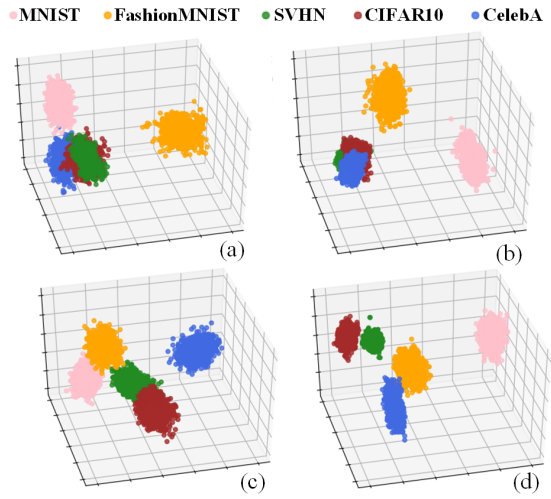

Figure 7: Feature visualization of 1000 samples from CIFAR10, extracted from the uncertainty estimation space (Z). The model is trained with $\mathcal{L}_e$ (b), $\mathcal{L}_r$ (c) and $\mathcal{L}_{total}$ (d), with center image-erasing. The control experiment is training the model with $\mathcal{L}_{total}$, without image-erasing.

ID	OOD	Ours		Ours ($\mathcal{H}$)	
		AUROC	AUPR	AUROC	AUPR
C10	C100	64.1	52.3	50.1	49.4
C100	C10	62.2	53.4	48.4	49.8

Table 3: The OOD detection results on CIFAR10 (C10) and CIFAR100 (C100), the group size is 10.

performance in Table 3 is that the current image-erasing strategy cannot create exclusive conditional entropy distribution for CIFAR10 and CIFAR100. The performance improvement compared with the detection results of real conditional entropy proves that our model has the ability to capture conditional entropy. An adaptive image-erasing strategy can be further investigated to address the limitation.

Conclusion

We proposed a method to perform transferable OOD detection by leveraging the concept of conditional entropy to OOD detection. We first validated two hypotheses: The DGMs are prone to learn low-level features rather than semantic information. In the DGMs, the lower bound of negative-log-likelihoods is determined by the conditional entropy between the model input and target output. Based on these hypotheses, we presented an image-erasing strategy and UEN to assign and capture the conditional entropy distribution discrepancy between different ID datasets. Our model, trained on a complex dataset, becomes transferable to other ID datasets. Experimental results on the five datasets show that our method, without retraining, achieves comparable performance with the SOTA group-based OOD detection methods that require retraining on the ID datasets.

References

BBC. 2020. AI "Outperforms" Doctors Diagnosing Breast Cancer. https://www.bbc.com/news/health-50857759. Accessed: 2020-01-02.

Bevandic, P.; Kreso, I.; Orsic, M.; and Segvic, S. 2018. Discriminative out-of-distribution detection for semantic segmentation. *CoRR*.

Cai, M.; and Li, Y. 2023. Out-of-distribution Detection via Frequency-regularized Generative Models. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Casas, S.; Sadat, A.; and Urtasun, R. 2021. MP3: A Unified Model To Map, Perceive, Predict and Plan. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *International Conference on Machine Learning (ICML)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Djurisic, A.; Bozanic, N.; Ashok, A.; and Liu, R. 2023. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations (ICLR)*.

Hsu, Y.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, H.; Li, Z.; Wang, L.; Chen, S.; Zhou, X.; and Dong, B. 2021. Feature Space Singularity for Out-of-Distribution Detection. In *Proceedings of the Workshop on Artificial Intelligence*.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiang, D.; Sun, S.; and Yu, Y. 2022. Revisiting flow generative models for Out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*.

Kim, E.; Kim, S.; Seo, M.; and Yoon, S. 2021. XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why Normalizing Flows Fail to Detect Out-of-Distribution Data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018a. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *International Conference on Learning Representations (ICLR)*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018b. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *International Conference on Learning Representations (ICLR)*.

Mohseni, S.; Pitale, M.; Yadawa, J. B. S.; and Wang, Z. 2020. Self-Supervised Learning for Generalizable Out-of-Distribution Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Nalisnick, E. T.; Matsukawa, A.; Teh, Y. W.; and Lakshminarayanan, B. 2019. Detecting Out-of-Distribution Inputs to Deep Generative Models Using a Test for Typicality. *CoRR*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *Workshop on Deep Learning and Unsupervised Feature Learning (NeurIPS workshop)*.

Nguyen, A. M.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nguyen, H. N.; Hung-Quang, N.; Ta, T.; Nguyen-Tang, T.; Doan, K. D.; and Thanh-Tung, H. 2023. A Cosine Similarity-based Method for Out-of-Distribution Detection. *CoRR*.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; DePristo, M. A.; Dillon, J. V.; and Lakshminarayanan, B. 2019. Likelihood Ratios for Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Salimans, T.; Karpathy, A.; Chen, X.; and Kingma, D. P. 2017. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifica-

tions. In *International Conference on Learning Representations (ICLR)*.

Sensoy, M.; Kaplan, L. M.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J. F.; and Luque, J. 2020. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models. In *International Conference on Learning Representations (ICLR)*.

Shekhovtsov, A.; and Flach, B. 2019. Feed-forward Propagation in Probabilistic Neural Networks with Categorical and Max Layers. In *International Conference on Learning Representations (ICLR)*.

Sun, R.; Zhang, A.; Zhang, H.; Zhu, Y.; Zhang, R.; and Li, Z. 2023. SR-OOD: Out-of-Distribution Detection via Sample Repairing. *CoRR*.

Times, T. N. Y. 2018. After Fatal Uber Crash, a Self-Driving Start-Up Moves Forward. https://www.nytimes.com/2018/05/07/technology/uber-crash-autonomous-driveai.html. Accessed: 2018-05-07.

Tishby, N.; Pereira, F. C. N.; and Bialek, W. 2000. The information bottleneck method. *CoRR*.

Wang, H.; Shi, X.; and Yeung, D. 2016. Natural-Parameter Networks: A Class of Probabilistic Neural Networks. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*.

Xiao, Z.; Yan, Q.; and Amit, Y. 2020. Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zaeemzadeh, A.; Bisagno, N.; Sambugaro, Z.; Conci, N.; Rahnavard, N.; and Shah, M. 2021. Out-of-Distribution Detection Using Union of 1-Dimensional Subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, L. H.; Goldstein, M.; and Ranganath, R. 2021. Understanding Failures in Out-of-Distribution Detection with Deep Generative Models. In *International Conference on Machine Learning (ICML)*.

Zhang, Y.; Liu, W.; Chen, Z.; Wang, J.; Liu, Z.; Li, K.; Wei, H.; and Chen, Z. 2020. Out-of-Distribution Detection with Distance Guarantee in Deep Generative Models. *CoRR*.