DMR: Disentangling Marginal Representations for Out-of-Distribution Detection

Dasol Choi*
Yonsei University, MODULABS
dasolchoi@yonsei.ac.kr

Dongbin Na*†
Pohang University of Science and Technology
dongbinna@postech.ac.kr

Abstract

Out-of-Distribution (OOD) detection is crucial for the reliable deployment of deep-learning applications. When a given input image does not belong to any categories of the deployed classification model, the classification model is expected to alert the user that the predicted outputs might be unreliable. Recent studies have shown that utilizing a large amount of explicit OOD training data is helpful for improving OOD detection performance. However, collecting explicit real-world OOD data is burdensome, and pre-defining all out-of-distribution labels is fundamentally difficult. In this work, we present a novel method, ***Disentangling Marginal Representations (DMR)****, that generates artificial OOD training data by extracting marginal features from images of an In-Distribution (ID) training dataset and manipulating these extracted marginal representations. DMR is intuitive and can be used as a realistic solution that does not require any extra real-world OOD data. Moreover, our method can be simply applied to pre-trained classifier networks without affecting the original classification performance. We demonstrate that a shallow rejection network that is trained on the small subset of synthesized OOD training data generated from our method and attachable to the classifier network achieves superior OOD detection performance. With extensive experiments, we show that our proposed method significantly outperforms the state-of-the-art OOD detection methods on the broadly used CIFAR-10 and CIFAR-100 detection benchmark datasets. We also demonstrate that our proposed method can be further improved when combined with existing methods. The source codes are publicly available at* https://github.com/ndb796/DMR.

1. Introduction

Deep-learning applications have produced remarkable success in a variety of domains, especially in image recognition tasks [5, 21, 25, 30, 33]. Meanwhile, the trustworthiness of predicted outputs from a trained model has also been required unprecedentedly for the reliable deployment of deep-learning applications [7, 19, 29]. For example, an autonomous driving vehicle utilizing deep neural networks could need to distinguish unknown objects that do not belong to any categories used in the training dataset. Out-of-Distribution (OOD) detection, thus, is a crucial tool to notify the users who use the deep-learning models of a degree of reliability for the predicted results. For example, input data might be rejected when the confidence score (probability) of the predicted output is below a certain threshold value. In the real-world deployment setting, a classification model can bump into not only known objects from the in-distribution $\mathcal{D}^{ID}$ but also unknown objects from the out-of-distribution $\mathcal{D}^{OOD}$. Therefore, the purpose of OOD detection is to train a rejection network $R(\cdot)$ that informs users of whether an input data x belongs to the $\mathcal{D}^{OOD}$ as follows:

$$R(x) = \begin{cases} 0 & \text{if } x \in \mathcal{D}^{ID} \\ 1 & \text{if } x \in \mathcal{D}^{OOD} \end{cases} \quad (1)$$

To implement the classifier $R(\cdot)$ that rejects the potential OOD data, we can (1) utilize the output representations from the pre-trained classification model itself in the inference time [2, 7, 31] or (2) train the rejection network on the pre-defined OOD dataset in the training time [8, 14, 17]. Recently proposed post-hoc OOD detection methods utilize feature vectors or logits of pre-trained classification models [2, 7, 18–20]. For example, the maximum value of the softmax output probabilities (MSP) can be used for the indicator that represents the probability of whether a given data is ID or OOD data [7, 19]. Generally, we expect that the MSP value of a classifier model trained on an ID training dataset is higher when a given data belongs to the ID data distribution compared to the OOD data distribution. On

†Correspondence to dongbinna@postech.ac.kr
*These authors contributed equally to this work.

the other hand, some methods leverage pre-defined training OOD data in the training time. Outlier Exposure (OE) enhances OOD detection performance by collecting a large number of explicit OOD data and using these data as OOD training data in the training time [8].

Some studies have adopted OOD data generation methods that synthesize OOD training data by utilizing *only* the ID training dataset and train a rejection network on the synthesized OOD training data. For example, Generative Adversarial Networks (GAN) can be used for generating synthesized OOD data that is classified as a uniform distribution by a classification model [17]. In their work, they argue that the OOD training data sufficiently close to the original in-distribution dataset is useful for training rejection networks. A recent work, KIRBY [14], has proposed an OOD data generation pipeline, which is motivated by the previous work [17]. They have demonstrated the images in which key regions are removed can be used as useful artificial OOD data for training a rejection network and their method shows improved OOD detection performance. However, the main idea of KIRBY is to aim to remove only the key regions, thus, representations of the background in an image still remain. Therefore, KIRBY might not be useful in the case where a large number of the data have no clear key regions.

Our work is also motivated by the useful observation [17] that the generated OOD training images close to the original in-distribution dataset are helpful in training a rejection network. However, in our study, *surprisingly*, we have observed that the OOD data does not need to be close to the in-distribution dataset in human perception. Our method, DMR, generates the artificial OOD training data by extracting marginal features from given images and combining the pieces of marginal representations. Moreover, our method can be applied to images that have no dominant key features. The generated OOD training images often seem unrealistic, however, greatly useful for training the rejection network. Our proposed method is an intuitive and novel approach that separates non-class discriminative marginal features from the ID dataset without explicitly selecting the key regions. We have demonstrated that our proposed method achieves superior OOD detection performance compared to the recent SOTA methods including generation-based methods [3, 6, 14, 17] and is even competitive with the OE, although our method does not require any extra real-world training OOD dataset and utilizes only the original ID dataset. Furthermore, we have demonstrated the OOD detection performance additionally increases by combining our method and the previous work. We hope that our findings provide new insight into the OOD detection research domain. We also provide the source codes.

2. Related Work

Related OOD detection studies could be divided into two categories (1) *post-hoc* methods and (2) *training* methods that utilize the OOD training data. First, recent studies have proposed various post-hoc OOD detection methods that are post-attachable given a pre-trained classifier. Post-hoc OOD detection methods generally utilize feature vectors or logits of pre-trained classification models [2, 7, 9, 19, 20, 28, 31]. These post-hoc methods do not require additional model training, thus, we can apply the post-hoc method to models that are already trained on a specific dataset. However, some post-hoc methods require additional inference time due to the gradient calculation [18, 19] or feature processing [31]. Secondly, we can train a network on the pre-defined OOD dataset [8, 14, 17] in the training time. OE is one of the most straightforward approaches, which aims to collect a large number of OOD data that does not belong to the ID categories and train a rejection network on this pre-defined OOD training dataset [8]. The OE method shows superior OOD detection performance compared to most post-hoc methods. However, the OE suffers from large labeling costs because they require an explicitly defined OOD training dataset. Unfortunately, the manifold of the true OOD data distribution is too large and unknown, therefore, OE is not feasible in many real-world deployment settings.

Alternatively, we can synthesize artificial OOD data and train the rejection network on the synthesized OOD dataset [13, 14, 17]. KIRBY [14] and GAN [17] only require the ID training dataset, which is a feasible setup for real-world deployment scenarios. The GAN can be used to generate OOD training images that are classified as a uniform distribution for a classifier [17]. However, their method needs to jointly train the three networks, which introduces training instability. Moreover, some synthesized OOD images with poor fidelity might result in OOD detection performance degradation [14]. In this perspective, the KIRBY that erases the key regions of an image utilizing an off-the-shelf pre-trained classifier [14] could be a better choice. However, we have found that KIRBY can not completely remove key regions of images, thus some synthesized OOD images are still class-discriminative. In contrast, our work takes a new direction, that extracts various marginal representations in a latent space and then synthesizes artificial OOD data by leveraging these marginal features. We have found that a simple latent vector manipulation can take off the marginal features from a set of feature representations of an ID dataset in the latent space. Our proposed method is novel and quite different from the recently proposed synthesizing methods [14, 17]. The previously proposed KIRBY is a top-down approach that detects the location of key regions in an ID training image and then erases the key features from that original image step

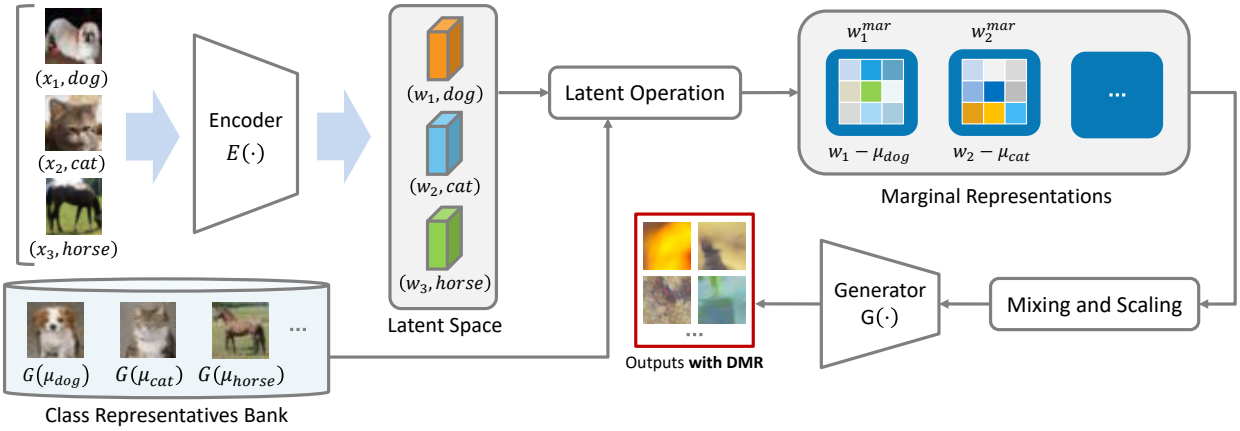

Figure 1. The illustration of our proposed method, DMR, which disentangles marginal features from the original image by leveraging the latent operation in the latent space. As shown in the above figure, our proposed method conducts the multiple latent mix-up (MLM) using the marginal representations w^{mar} to generate the artificial OOD training images. Our feature manipulation method in the latent space is effective in generating useful OOD training images.

by step. In contrast, our method takes a novel bottom-up approach that combines various pieces of marginal representations by leveraging the latent vectors. We also find a new observation through the synthesized OOD training images generated from our method, that our generated OOD training images *are sometimes not a natural image in human perception, however greatly effective for training the OOD rejection network*, which has not been observed previously. Our method can also be combined with another synthesizing method such as KIRBY, which further improves the OOD detection performance and indicates that our approach can be used as an orthogonal method.

3. Proposed Methods

For classification tasks, we train a classification model on training samples of the in-distribution $\mathcal{D}^{ID}$ over pairs of data x and corresponding labels y. We generally train a model containing a feature extractor $F\colon \mathcal{X} \to \mathcal{Z}$ and a shallow classifier $C\colon \mathcal{Z} \to \mathcal{Y}$ by minimizing the empirical risk:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}^{ID}}[\ell(C(F(x)), y)] \tag{2}$$

where ℓ denotes a suitable loss function such as the cross-entropy function for a multi-class classification task, $\mathcal{X}$ denotes input space, $\mathcal{Y}$ denotes a label space, and $\mathcal{Z}$ denotes a feature space. Our method aims to generate the artificial training OOD images that compose the proxy OOD distribution $\mathcal{D}^{OOD}_{train}$. Our proposed method produces this OOD training dataset using *only* samples of the in-distribution $\mathcal{D}^{ID}$ because we assume that we can not access the true test OOD distribution $\mathcal{D}^{OOD}_{test}$ in the training time.

To generate the synthesized OOD dataset, we utilize individual two functions, encoder $E\colon \mathcal{X} \to \mathcal{W}$ and generator $G\colon \mathcal{W} \to \mathcal{X}$ where $\mathcal{W}$ denotes a latent space. We formally introduce the ideal disentangling property of a generator G is the ability to separate the latent space $\mathcal{W}$ into $\mathcal{W}^{cls}$ and $\mathcal{W}^{mar}$. We hypothesize that $\mathcal{W}^{cls}$ consists of the class-discriminative features that are distinguished from the marginal representation distribution $\mathcal{W}^{mar}$. Our method utilizes encoded latent vectors $w^{ori} = (w^{mar}, w^{cls})$ with $w^{mar} \in \mathcal{W}^{mar}$ and $w^{cls} \in \mathcal{W}^{cls}$. For a given image x, the encoded latent vector w^{ori} can be obtained by forwarding the original sample x into an encoding network $E(\cdot)$ where the $E(\cdot)$ maps image x into a latent space of $G(\cdot)$ [1, 26]. Ideally, the original sample x can be reconstructed as $G(w^{mar}, w^{cls})$ using a generative model $G(\cdot)$.

We note that the marginal representations w^{mar} are also important to generate valid images. For example, some textures that do not contain class-discriminative features can compose a feasible image. The background scenes such as clouds and sky can be found in various images along diverse categories such as ocean, glacier, and forest for scene classification tasks, which might be interpreted as being not class-discriminative. Using these ingredients, we could train the rejection network $R(\cdot)$ to classify inputs as follows:

$$R(x) = \begin{cases} 0 & \text{if } x = G(w^{ori}) \\ 1 & \text{if } x = G(w^{mar}) \end{cases} \tag{3}$$

In this work, we regard the rejection network as $R = F \cdot B$ where $F(\cdot)$ is the frozen feature extractor trained on the $\mathcal{D}^{ID}$ and the $B(\cdot)$ denotes the additional binary classification model. For OOD detection, our proposed method only trains the model $B(\cdot)$ that is a shallow MLP attachable to the frozen feature extractor $F(\cdot)$.

KIRBY
(AAAI 2023)

Multiple Latent Mix-up
without DMR (Ours)

Multiple Latent Mix-up
with DMR (Ours)

Figure 2. The synthesized OOD training data examples are randomly selected from each method using the CIFAR-10 ID dataset. With DMR, the class-discriminative features are relatively well erased from the original images compared to the KIRBY and the vanilla MLM.

3.1. Multiple Latent Mixup (MLM)

Our work focuses on the way to synthesize artificial OOD training data. First, we start with an assumption of previous work. A study [17] argues that the synthesized OOD training samples nearby in-distribution samples are useful to train the rejection network and they utilize a GAN architecture to generate images that are classified as a uniform distribution. To meet this property, we first present *Multiple Latent Mix-up (MLM)* which mixes the encoded latent vectors in the latent space.

$$x_{mix}^{OOD} = G(\frac{1}{k}\sum_{i}^{k} w_i^{ori}) \approx G(w^{mar}) \tag{4}$$

We can set the number of samples to mix by adjusting the value of k. We recommend setting the classes of mixed samples to be different from each other to remove representative features belonging to certain categories. For example, if $k = 1$, the decoded image $G(w^{ori})$ is ideally the same as the original image x. In contrast, if the k is sufficiently large, the diversity of the synthesized increases, which results in more abundant OOD training data. However, if the $k \approx M$ where M denotes the number of classes in the training dataset, the sampled diversity drastically decreases. We have found that OOD training images generated from our vanilla MLM are more effective than the previous work [17] to train the rejection network, however, some synthesized images are highly correlated with a certain class and still contain class-discriminative features as shown in Figure 2.

3.2. Disentangling Marginal Representations

To remedy the limitation of vanilla MLM, we aim to synthesize OOD data that does not contain class-correlated representations. We postulate that a latent representation w^{ori} can be conceptually divided into (1) class-discriminative representations w^{cls} that have a high correlation with a true label y and (2) the marginal representations w^{mar} whose features are not correlated to a certain class. To extract marginal representations from data, we leverage the class-representative features. In the latent space, we first define the representative latent vectors as $\mu_1, \mu_2, \ldots, \mu_M$ for each class where M is the total number of classes. We expect that the latent representatives μ_y contain class-discriminative features according to their true class y by taking the average latent representations per class. Thus, we can extract the marginal representations by utilizing a simple latent manipulation given an in-distribution image data x and corresponding label y. In this work, our proposed method obtains the marginal representations by utilizing latent vectors as follows:

$$w^{mar} = E(x) - \mu_y \tag{5}$$

where $\mu_y = \mathbb{E}_{x \in \mathcal{D}_y^{ID}}[E(x)]$ and $\mathcal{D}_y^{ID}$ denotes the original in-distribution image dataset that belongs to the label y. Our method extracts the marginal features of x by subtracting the μ_y from $E(x)$. Interestingly, we have observed this simple latent operation is effective in obtaining useful OOD training samples. To generate high-fidelity OOD training samples, we adopt the recently proposed style-based GAN model [12, 27]. We note that the encoding procedure is optional because we can get simply random latent vector w from the GAN model [10, 11]. For example, we might use the mapping network [10] instead of encoding an image sample into a latent vector. With the assumption that the images generated from a generative model $G(\cdot)$ are on the valid image manifold, we can obtain the synthesized OOD data by removing class-discriminative features using the following equation (single latent inversion **with DMR**):

$$x_{DMR}^{OOD} = G(\lambda \cdot w^{mar}) = G(\lambda \cdot (E(x) - \mu_y)) \tag{6}$$

ID	Methods	OOD Datasets SVHN AUROC ↑	SVHN AUPR ↑	Textures AUROC ↑	Textures AUPR ↑	LSUN-crop AUROC ↑	LSUN-crop AUPR ↑	Tiny-ImageNet AUROC ↑	Tiny-ImageNet AUPR ↑	Place-365 AUROC ↑	Place-365 AUPR ↑	Gaussian Noise AUROC ↑	Gaussian Noise AUPR ↑	Average AUROC ↑	Average AUPR ↑
CIFAR-10	MSP	91.91	95.81	88.51	78.50	96.49	95.69	94.59	93.10	88.24	95.61	83.69	73.73	90.57	88.74
	ODIN	91.63	95.95	88.34	80.70	97.49	97.31	95.80	95.19	87.92	95.69	75.54	63.94	89.45	88.13
	Mahalanobis	96.78	98.72	96.26	94.55	93.73	92.33	86.90	82.20	80.15	92.33	99.23	98.24	92.18	93.06
	Energy	91.07	96.03	85.34	78.47	99.05	99.02	97.97	97.75	89.88	96.76	62.63	53.91	87.66	86.99
	Entropy	92.39	96.31	88.83	80.38	97.32	97.13	95.43	94.79	88.70	96.03	83.09	72.59	90.96	89.54
	MaxLogit	91.10	96.07	85.48	78.65	98.95	98.89	97.83	97.56	89.81	96.71	65.46	56.20	88.11	87.34
	KL-Matching	84.34	92.50	75.98	69.80	90.83	90.97	83.62	86.42	72.07	91.58	55.80	56.50	77.11	81.30
	ViM	95.58	98.16	93.78	90.89	98.48	98.45	95.82	95.40	85.60	94.81	95.26	92.04	94.09	94.96
	GAN	72.40	74.82	75.33	72.65	72.64	72.19	76.81	74.75	81.03	92.55	69.45	57.56	74.61	74.08
	ACET	91.85	96.22	88.74	79.01	92.56	90.97	88.21	88.93	89.02	95.61	86.04	80.11	89.40	88.48
	MIM	97.25	98.58	**96.11**	**94.13**	99.52	99.41	**98.72**	**98.55**	90.94	96.90	99.65	99.25	97.03	97.80
	KIRBY	99.03	99.63	92.26	90.67	99.55	99.51	97.93	97.85	89.86	96.74	99.05	98.02	96.28	97.07
	Ours (DMR)	**99.46**	**99.79**	95.81	94.03	**99.59**	**99.55**	98.16	98.02	**93.29**	**97.86**	**99.70**	**99.29**	**97.67**	**98.09**

Table 1. Comparison with state-of-the-art methods using a WideResNet-40-2 classifier. All experiments are conducted by the OOD detection benchmark framework [15]. The symbol ↑ indicates larger values are better.

where the λ denotes an emphasis factor for amplifying marginal representations. We generate $\mathcal{D}^{OOD}_{DMR}$ dataset using the above equation. For experiments, we simply choose the constant λ as a real number that is uniformly sampled between 1 and 3, which is suitable to obtain the improved OOD detection performance. While the generated OOD data x^{OOD}_{DMR} sometimes does not seem realistic in human perception, we demonstrate that these images can be greatly helpful for training the rejection network.

However, we have observed that the simple subtraction of latent vectors might suffer from the lack of diversity of synthesized images. Thus, we further leverage the Multiple Latent Mixup (MLM) that mixes representations in the latent space with our proposed DMR. In conclusion, the artificial OOD training data $x^{OOD}_{DMR\_mix}$ can be generated by the following equation given an original image x and corresponding label y (multiple latent mix-up **with DMR**):

$$x^{OOD}_{DMR\_mix} = G(\lambda \cdot \frac{1}{k}\sum_{i}^{k} w_i^{mar}) \tag{7}$$

We generate a $\mathcal{D}^{OOD}_{DMR\_mix}$ dataset using the above equation. The synthesized OOD training data based on our methods are illustrated in Figure 2. Our DMR generates artificial OOD images utilizing the marginal representations while the KIRBY [14] removes the key regions that generally contain class-discriminative objects. We note that our proposed method can be applied to images that have no clear class-discriminative key regions.

3.3. Training Networks

We first train various neural networks, WideResNet-40-2 (WRN), ResNet-50, and DenseNet-121 on the CIFAR-10 and the CIFAR-100 datasets individually. We then freeze these trained classification models. Given a frozen feature extractor $F(\cdot)$ of a classifier, we train only an additional binary classification rejection network $B(\cdot)$ which is a 2-layer shallow MLP network on the synthesized OOD dataset from our proposed method. After training, we can simply detect the OOD samples by forwarding the feature vectors extracted by the frozen feature extractor $F(\cdot)$ into the shallow rejection network $B(\cdot)$. Our proposed method does not update the weights of the original classification model $F(\cdot)$ and the classification head $C(\cdot)$, thus, does not affect the original classification performance. Our final loss function is based on the binary cross-entropy function as follows:

$$-\mathbb{E}[y \cdot log(R(x)) + (1-y) \cdot log(1-R(x))] \tag{8}$$

where we set $x^{ID} \in \mathcal{D}^{ID}$ to a negative sample ($y=0$) and synthesized x^{OOD}_{train} to a positive sample ($y=1$). In detail, the $\mathcal{X}^{OOD}_{train}$ is determined as described in Table 3. For removing the class-discriminative features from the original ID data, our proposed method takes a surrogate approach that generates x^{OOD}_{train} samples that do not belong to a specific category. Our proposed method is an effective and realistic solution in that our method can be applied to any off-the-shelf frozen classifier and does not affect the original classification performance.

3.4. Latent Manipulation

For the experiments, we have adopted the recently presented StyleGAN architecture [27] which has been known as being well-disentangling feature representations. Specifically, we train the StyleGAN-XL model where the dimension of the latent vectors is 12×512 on the CIFAR-10 dataset. We have also trained the StyleGAN-XL model where the dimension of the latent vectors is 32×512 on the ImageNet-1k dataset. Recent studies also have demonstrated that the

ID	Methods	OOD Datasets													
		SVHN		Textures		LSUN-crop		Tiny-ImageNet		Place-365		Gaussian Noise		Average	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
CIFAR-100	OE	93.53	96.53	86.84	77.33	91.28	89.68	91.96	90.47	**84.59**	**94.35**	94.85	86.35	90.50	89.11
	KIRBY	94.35	95.53	89.08	85.70	**97.11**	**96.36**	91.91	90.81	78.49	92.10	99.92	99.83	91.81	93.38
	Ours (DMR)	**98.81**	**99.51**	89.39	84.59	9628	95.40	**92.88**	90.65	77.75	91.71	99.94	99.69	92.50	93.59
	Ours + KIRBY	98.39	99.33	**91.65**	**87.75**	96.90	96.15	92.72	**90.92**	78.71	92.10	**99.95**	**99.74**	**93.05**	**94.33**

Table 2. Comparison results with OE. Without any real-world OOD training data, ours shows a better performance than the OE.

encoding (inversion) method could be used for finding the corresponding latent vector given image x [1, 26]. First, we utilize the encoding method to map an image x into the latent vector w^{ori}. Previous studies have shown that latent vector arithmetic is effective for semantic manipulation in the image generation domain [1, 23]. However, we have found that the simple subtraction in the latent space may suffer from the diversity of generated images. That is why we further utilize the multiple mix-ups (MLM) in the latent space, which is a novel and effective approach.

MLM	DMR	OOD Trainset $\mathcal{X}^{OOD}_{train}$	CIFAR-10 AUROC	CIFAR-10 AUPR	CIFAR-100 AUROC	CIFAR-100 AUPR
✓		$\mathcal{D}^{OOD}_{mix}$	95.27	96.26	89.80	92.52
	✓	$\mathcal{D}^{OOD}_{DMR}$	97.65	98.01	91.80	92.81
✓	✓	$\mathcal{D}^{OOD}_{DMR\_mix}$	**97.67**	**98.09**	**92.50**	**93.59**

Table 3. OOD detection performances of proposed methods averaged over the six OOD benchmark datasets.

4. Experiments

4.1. Datasets

For the experiments, we follow the broadly adopted OOD detection evaluation settings used in previous studies [14, 15]. We use CIFAR-10 and CIFAR-100 [16] as the in-distribution datasets, which are widely utilized in numerous image recognition tasks. For training the attachable rejection network, we generate 50,000 OOD images for the CIFAR-10 classifier and the CIFAR-100 classifier. For example, we train the rejection network $B(\cdot)$ on the 50,000 original ID training images and synthesized OOD dataset $\mathcal{D}^{OOD}_{train}$ whose size is 50,000, while freezing the feature extractor $F(\cdot)$ of a classifier trained on the CIAFR-10 dataset.

4.2. Evaluation Metrics

We have adopted the baseline evaluation metrics in the OOD detection research domains as follows:

- **AUROC** is a generally adopted metric evaluating the detection performance of a binary classifier, which denotes the area under the receiver operating characteristics.
- **AUPR** denotes the area under the precision-recall curve. This metric is especially useful for evaluating binary classification performance in imbalanced data settings.

4.2.1 ID and OOD Datasets

We experiment with the proposed methods using the CIFAR-10 and the CIFAR-100 dataset as the in-distribution dataset. We use the following OOD test datasets for evaluating various OOD detection methods identically.

- **SVHN** consists of images that represent a digit between 0 and 9. The SVHN has been used for OOD detection generally in the case that the ID dataset is CIFAR-10 or CIFAR-100 [24].
- **Textures** contains diverse images belonging to natural texture categories. This dataset is also known as Describable Texture Dataset (DTD) and has 47 categories. Each category has 120 images [4].
- **LSUN-crop** dataset is devised for scene classification tasks, containing various scene images. The categories of LSUN are not overlapped with the CIFAR-10 and CIFAR-100 dataset [32].
- **Tiny-ImageNet** dataset contains a small subset of the original ImageNet dataset. In detail, we select the Tiny-ImageNet cropped dataset as a test OOD dataset following previous studies [29].
- **Places-365** dataset contains various images and each image belongs to a certain place (scene) over 365 classes [22].
- **Gaussian Noise** dataset contains random noise images sampled from the Gaussian distribution.

4.3. Overall Experimental Results

We first reproduce the recently proposed SOTA methods and report their OOD detection performance in Table 1 and Table 4.

For the CIFAR-10 and CIFAR-100 datasets, we train the rejection network $B(\cdot)$ using all 50,000 $\mathcal{D}^{ID}$ images and similarly synthesized 50,000 $\mathcal{D}^{OOD}_{train}$ images for both KIRBY and our method. The rejection network $B(\cdot)$ is trained for 10 epochs, with a learning rate of 0.01 and a momentum of 0.9. We use a mix-up parameter $k = 5$ and

ID	Methods	OOD Datasets													
		SVHN		Textures		LSUN-crop		Tiny-ImageNet		Place-365		Gaussian Noise		Average	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
CIFAR-100	MSP	71.37	84.37	73.54	57.50	85.58	84.35	86.32	84.80	73.91	89.44	80.67	69.09	78.56	78.25
	ODIN	64.69	78.65	72.61	57.29	85.69	84.68	87.36	86.23	73.08	88.93	77.96	65.48	76.89	76.87
	Mahalanobis	85.68	93.09	89.92	85.28	52.03	46.74	55.95	49.22	63.92	84.24	99.90	99.87	74.56	76.40
	Energy	48.28	69.66	53.53	39.70	96.91	97.12	95.92	95.81	63.49	84.32	70.40	56.85	71.42	73.74
	Entropy	73.87	85.01	76.29	61.57	95.88	95.71	95.28	94.92	75.81	90.42	79.52	66.08	82.77	82.28
	MaxLogit	73.95	85.36	76.37	61.64	95.14	94.68	94.67	94.01	75.94	90.50	72.37	58.60	81.40	80.79
	KL-Matching	70.14	86.76	71.89	63.10	77.72	80.41	82.09	81.80	66.06	88.07	81.27	74.09	74.86	79.03
	ViM	92.54	96.64	90.69	86.26	75.55	65.73	79.96	71.54	70.11	87.68	99.88	99.84	84.78	84.61
	MIM	88.92	94.35	88.02	83.92	95.04	94.68	92.36	**91.17**	71.91	87.59	99.79	**99.96**	89.34	91.94
	KIRBY	94.35	95.53	89.08	**85.70**	**97.11**	**96.36**	91.91	90.81	**78.49**	**92.10**	99.92	99.83	91.81	93.38
	DMR	**98.81**	**99.51**	**89.39**	84.59	96.28	95.40	**92.88**	90.65	77.75	91.71	**99.94**	99.69	**92.50**	**93.59**

Table 4. Comparison with state-of-the-art methods using a WideResNet-40-2 classifier on CIFAR-100. The symbol ↑ indicates that larger values are better.

ID	CIFAR-10			CIFAR-100		
	WRN	ResNet	DenseNet	WRN	ResNet	DenseNet
MSP	90.57	84.96	85.24	78.56	67.51	76.02
ODIN	89.45	84.30	86.18	76.89	67.02	75.22
Mahalanobis	92.18	85.11	85.71	71.42	57.50	80.52
Energy	87.66	83.08	83.10	76.98	75.86	83.31
Entropy	90.96	85.28	85.78	82.77	70.43	79.49
MaxLogit	88.11	83.79	84.13	81.40	74.88	83.38
KL-Matching	77.11	74.25	75.95	74.86	65.19	68.40
ViM	94.09	91.57	92.32	84.78	83.09	83.34
MIM	97.17	96.97	97.11	89.37	82.33	81.01
KIRBY	96.26	97.08	97.18	91.81	81.12	82.90
Ours (DMR)	**97.67**	**97.36**	**97.24**	**92.50**	**83.53**	**83.51**

Table 5. Experimental results (AUROC) using different architectures averaged over the six OOD benchmark datasets.

an emphasis factor λ sampled from a uniform distribution between 1 and 3.[1]

Detailed examples of synthesized OOD samples according to the mix-up parameter k and emphasis factor λ are illustrated in Figure 3 and Figure 4. Interestingly, our proposed method shows better performance over the all OOD test datasets compared to the KIRBY that generates background-like images. This result indicates that the synthesized OOD training data does not need to be natural images in human perception. Although our synthesized OOD images are not realistic for humans, yet, are effective in training the rejection network.

Further detailed synthesized OOD data examples according to the various hyper-parameters k and λ are illustrated in the supplementary materials. Our methods show superior OOD detection performance compared to the other SOTA methods as shown in Table 1 and Table 4. Our proposed method can be combined with other synthesized methods such as KIRBY. Our experiments show the combined method achieves better performance than OE as shown in Table 2. We note that the OE [8] requires a large-scale explicit OOD training dataset beyond the ID dataset and also requires additional training steps. In contrast, our method only utilizes the ID training dataset. We have also shown our proposed method achieves superior OOD detection performance than SOTA methods over various network architectures as shown in Table 5.

[1]Code is available at https://github.com/ndb796/DMR

5. Discussion

A previous study has proposed the synthesizing method to generate OOD images that are classified as a uniform distribution [17]. However, this method shows poor detection performance due to the training instability introduced by their approach that jointly trains three networks including the GAN model. Our method utilizes the individually trained style-based GAN architectures [10, 11], which results in the high-fidelity samples synthesized OOD training data as shown in Figure 2. Using the multiple latent mix-ups **without DMR**, the generated OOD images seem more *realistic* and achieve a low FID score than the previous GAN-based method [17] in the datasets in this work.

In previous work KIRBY, they argue their improved detection performance is introduced by the close semantic distance between the ID and OOD datasets. However, we have found that the synthesized OOD data do not need to be close to the ID data distribution in human perception. Although the artificial OOD training images can seem unnatural for humans, these images can be largely useful for training the rejection network. Rather, we have found that if the ID dataset has too much semantic information similar to the ID dataset, training the rejection network might be hard. To verify this assumption, we have explored the OOD detection performance of our method that utilizes the synthesized images generated from the mix-up of latent vectors w^{ori} (MLM). As shown in Table 3, when using the synthesized OOD images generated by our vanilla multiple latent mix-

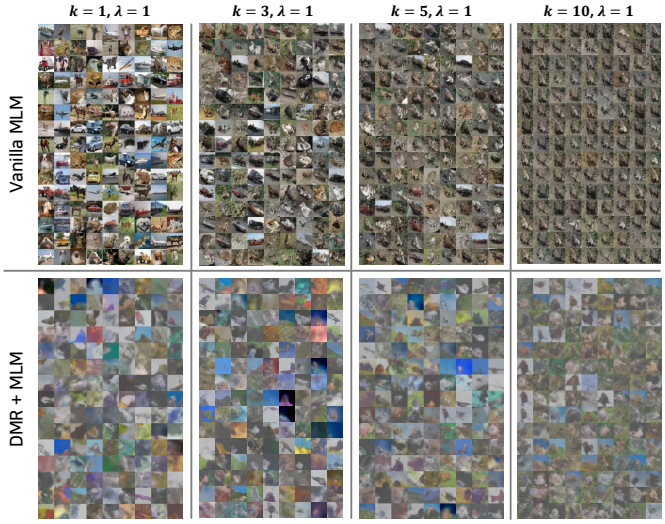

Figure 3. The examples of synthesized OOD images generated from our methods according to the value of the mixup parameter k using the CIFAR-10 dataset.

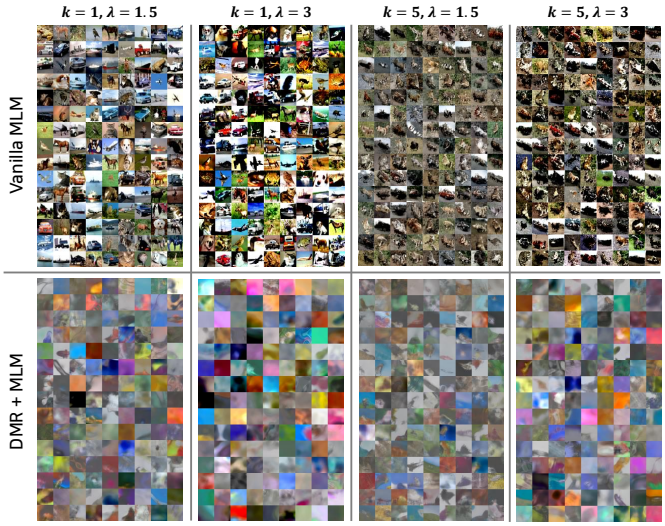

Figure 4. The examples of synthesized OOD images generated from our methods according to the value of the feature emphasis parameter λ using the CIFAR-10 dataset.

up method **without DMR**, the final OOD detection performance is relatively poor, which indicates it is also important to disentangle and remove the class-discriminative features. As shown in Figure 2, KIRBY erases the key features, however, KIRBY tends to leave some class-discriminative features and artifacts, thus frequently producing unsatisfactory synthesized samples. Our method is based on a recently presented GAN architecture, which results in better fidelity compared to the KIRBY. We have also demonstrated that it is important to remove the class-discriminative features while maintaining the synthesized OOD images to have good fidelity. Specifically, we have shown the OOD detection performance could be improved with our DMR (Table 3). The disentangling procedure is fundamentally hard for natural images and we believe the latent encoding method can be further improved. Thus, we will further explore suitable disentangling methods and latent inversion methods to generate artificial OOD data for future work.

We can control the degree of mixing and the level of the diversity of the synthesized OOD samples by adjusting the latent mix-up parameter k. When the k increases, the chance that we could obtain more diverse synthesized OOD images becomes larger. However, if the mix-up parameter k is the same as M ($k = M$) where M is the number of classes, the diversity of images drastically decreases. Therefore, we recommend setting $3 \leq k \leq 5$ for generating synthesized OOD images to achieve improved OOD detection performance for the CIFAR-10 ID setup. Our proposed method also utilizes an effective emphasis factor λ for scaling feature representations. We have observed that the diversity of images additionally increases by adjusting the λ values. As shown in Figure 4, the contrast and saturation of images tend to be enhanced when the value of the λ scaling factor increases, which could result in the improvement of the diversity of synthesized OOD training images. Thus, we adopt this emphasis approach for feature representations and we observe that this method further improves the OOD detection performance by generating useful OOD training samples. We also observe that if $3 \ll \lambda$, the brightness and contrast of synthesized OOD samples are enhanced too much to obtain a good OOD detection performance.

6. Conclusion

In this work, we present a novel method, DMR that disentangles the marginal representations from the original images. The synthesized OOD training data that combines these marginal features is greatly useful for OOD detection. The latent mix-up across the different categories can provide more diverse artificial samples. On the baseline OOD detection benchmark datasets, our proposed method shows superior performance compared to the recently proposed state-of-the-art methods. Our work revisits the desirable properties of synthesized OOD images and discusses them. We hope that this work provides new insights for OOD detection research.

7. Acknowledgements

This research was supported by Brian Impact, a non-profit organization dedicated to advancing science and technology. We also express our gratitude to MODULABS for their support and contributions to this research.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3, 6

[2] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021. 1, 2

[3] Dasol Choi and Dongbin Na. Towards reliable ai model deployments: Multiple input mixup for out-of-distribution detection. *arXiv preprint arXiv:2312.15514*, 2023. 2

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 2

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. 1, 2

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 1, 2, 7

[9] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4, 7

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4, 7

[12] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 4

[13] Jaeyoung Kim, Kyuheon Jung, Dongbin Na, Sion Jang, Eunbin Park, and Sungchul Choi. Pseudo outlier exposure for out-of-distribution detection using pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023. 2

[14] Jaeyoung Kim, Seo Taek Kong, Dongbin Na, and Kyu-Hwan Jung. Key feature replacement of in-distribution samples for out-of-distribution detection. 2023. 1, 2, 5, 6

[15] Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. Pytorch-ood: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4351–4360, 2022. 5, 6

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 6

[17] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 1, 2, 4, 7

[18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 2

[19] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1, 2

[20] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 1, 2

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[22] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020. 6

[23] Dongbin Na, Sangwoo Ji, and Jong Kim. Unrestricted black-box adversarial attack using gan with limited queries. In *ECCV 2022 Workshops*, 2023. 6

[24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3, 6

[27] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 4, 5

[28] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. 2

[29] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 1, 6

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[31] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 1, 2

[32] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC*, 2016. 1