Balanced Energy Regularization Loss for Out-of-distribution Detection

Hyunjun Choi[1],[2]* Hawook Jeong[2] Jin Young Choi[1]
[1] ASRI, ECE., Seoul National University [2] RideFlux Inc.
numb7315@snu.ac.kr hawook@rideflux.com jychoi@snu.ac.kr

Abstract

In the field of out-of-distribution (OOD) detection, a previous method that use auxiliary data as OOD data has shown promising performance. However, the method provides an equal loss to all auxiliary data to differentiate them from inliers. However, based on our observation, in various tasks, there is a general imbalance in the distribution of the auxiliary OOD data across classes. We propose a balanced energy regularization loss that is simple but generally effective for a variety of tasks. Our balanced energy regularization loss utilizes class-wise different prior probabilities for auxiliary data to address the class imbalance in OOD data. The main concept is to regularize auxiliary samples from majority classes, more heavily than those from minority classes. Our approach performs better for OOD detection in semantic segmentation, long-tailed image classification, and image classification than the prior energy regularization loss. Furthermore, our approach achieves state-of-the-art performance in two tasks: OOD detection in semantic segmentation and long-tailed image classification.

1. Introduction

Deep neural networks are used in a variety of fields such as image classification [22] and semantic segmentation [11]. However, there is a challenge in the practical use of deep neural networks in areas where safety is crucial, such as autonomous driving and medical diagnosis [20, 25]. In particular, deep neural networks have the issue of providing high confidence to out-of-distribution (OOD) samples that are not used for training [15]. As a result, Maximum softmax probability (MSP) score has been proposed to identify these OOD samples [17]. Based on the score, OOD detection performance is evaluated by metrics (e.g. AUROC, FPR). Both in image classification [18, 24, 26, 29, 30, 38, 40, 43, 46](including long-tailed image classification [43]) and semantic segmentation [1–3, 5, 10, 12, 16, 19, 28, 33, 36, 41], different approaches have been suggested to enhance the OOD detection performance. Among them, we concentrate on the methods using auxiliary data as OOD data which indicate superior OOD detection performance to the previous methods that only use in-distribution samples.

Outlier Exposure (OE) utilizes an auxiliary dataset of outliers to improve OOD detection performance [18]. The auxiliary data is consist of classes that do not overlap with the in-distribution data and the test OOD data. OE leverages the cross-entropy loss for the existing training data and the regularization loss for the auxiliary data. The cross-entropy loss that results from giving the auxiliary data a uniform label is the regularization loss of OE. Meanwhile, a new energy score has been introduced in Energy-based OOD detection (EnergyOE) which replaces the MSP score [29]. Furthermore, EnergyOE suggests an energy regularization loss that differs from that of OE to enhance performance. The squared hinge loss for energy with every existing (in-distribution) piece of data and every auxiliary (OOD) piece of data is added to create the energy regularization loss. Similarly, in semantic segmentation, the OOD detection performance is enhanced by using the auxiliary dataset of the outlier. Meta-OOD [5] organized the auxiliary dataset of the outlier by scenes of the COCO dataset [27]. Although the process of creating the auxiliary data is different from image classification, the training loss is comparable. Meta-OOD adopts the regularization loss proposed by OE. Recently, PEBAL [41] also adopts energy regularization loss proposed by EnergyOE.

However, when regularizing auxiliary data, the existing methods for OOD detection do not take into account variations between auxiliary data samples. The variations are severe especially on real data such as semantic segmentation for autonomous driving. As seen in Figure 1a, for the pretrained model, the class distribution of the auxiliary OOD data is not uniform across classes, i.e., imbalanced. To address the imbalanced problem, we regularize the auxiliary data differently for each sample. To achieve this, we propose a balanced energy regularization loss to apply higher regularization to majority classes than minority classes in

[*] Work done as an intern at RideFlux Inc.

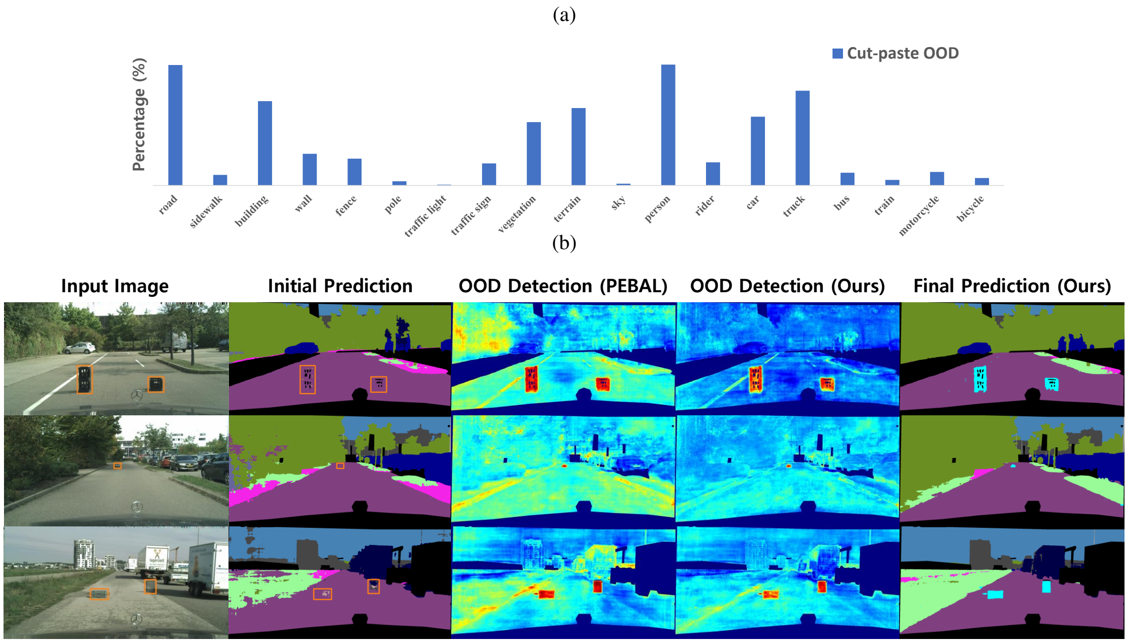


Figure 1. Overview of our approach in semantic segmentation task (a): Class distribution of cut-pasted OOD pixels collected from 10000 synthesized scene images ; (b): OOD detection result in Fishyscapes validation sets. Our balanced energy PEBAL(Ours) is the method that substitutes the energy regularization loss in PEBAL [41] with our balanced energy regularization loss.

auxiliary data. In other words, auxiliary samples of majority classes receive a larger energy constraint than samples of minority classes. We introduce the term Z, which indicates whether a sample belongs to the majority or minority of a class. Z is the weighted sum of the softmax output of the classification model for a sample (i.e., the posterior probability of a class for a given sample), where the weight is the prior probability for the class. Unlike the existing energy regularization loss, our balanced energy regularization loss adjusts to the value of Z for an auxiliary data sample. Two adaptive loss components make up our loss: loss margin and loss weight. The adaptive loss margin provides an additional Z-proportional margin in the squared hinge loss for auxiliary data. The adaptive loss weight gives a weight proportional to Z to the squared hinge loss.

We confirm our novel loss in three tasks: semantic segmentation, long-tailed image classification, and image classification. The proposed loss is simple but generally effective for various tasks. Figure 1b illustrates how our method outperforms the previous state-of-the-art (SOTA) algorithm PEBAL in the semantic segmentation task by replacing the energy regularization loss with our loss. OOD detection performance is also enhanced when using our loss compared to the baseline (EnergyOE) which use only energy regularization loss. In all image classification tasks, we evaluate our method on semantically coherent OOD detection (SC-OOD) benchmark [46]. In long-tailed image classification task, our approach reveals superior OOD performance compared to both OE and EnergyOE methods which use auxiliary data. In addition, our approach outperforms the previous SOTA method PASCL [43], Similarly, in the image classification task, we demonstrate the superiority of our loss by outperforming both OE and EnergyOE, which make use of auxiliary data. The contributions are summarized as:

- By making inferences based on previously trained models, we explain the imbalanced distribution of auxiliary OOD data.
- We suggest a novel balanced energy regularization loss to address the class imbalance in auxiliary OOD data.
- The proposed balanced loss performs better for OOD detection than the previous energy regularization loss.
- The SOTA performance for OOD detection in two tasks is achieved by our OOD detection method.

2. Related Work

2.1. OOD in Image Classification

In the image classification task, there are two main approaches for OOD detection in deep neural networks. The first is a method to attain prediction uncertainty for a pretrained model [24, 26, 29, 40]. The second approach modifies the architecture or loss based on new training [9, 15, 18, 29, 38]. The first method primarily suggests new measures to boost performance. Proposed measures include the baseline MSP [17], Mahalanobis distance [24], the distance from the distribution of the Gram matrix [40], and free energy whose value is computed from the logit rather than the probability [29].

The second method mainly uses auxiliary data as OOD data to enhance performance. Representatively, there are OE [18] and EnergyOE [29] as methods of using auxiliary data for learning. OECC [38] replaced OE's cross entropy loss with total variance loss and included calibration loss. In UDG [46], auxiliary data is once again divided into OOD samples and in-distribution samples through unsupervised dual grouping to improve performance. The more challenging SC-OOD benchmark is also proposed by UDG.

2.2. OOD in Semantic Segmentation

In general, semantic segmentation's OOD detection benchmarks [3, 16, 28] resolve the detection problem of OOD that appears in complex urban driving scenes [8]. The assessment method follows the same criteria as image classification [17], but instead of evaluating images as a whole, it does so pixel-by-pixel. Similarly, there are numerous ways to enhance OOD detection by putting forth new measures such as MSP [16], Entropy [17], Mahalanobis [24], and Energy [29]. Newly proposed methods include Max Logit [16], which is the maximum value of logit, and Standardized Max Logit (SML) [19], which is improved by reducing the deviation of logit by class. On the other hand, Bayesian Deeplab [36] measures uncertainty through a dropout layer based on Bayesian estimation. Image Resynthesis [28] and Synboost [10] are algorithms that reconstruct and use new data from existing data through a generative model.

Similar to the task of classifying images, the majority of the top techniques use auxiliary data as OOD data. The OOD data is synthesized by cut-pasting the object's mask from auxiliary dataset, or is raw pixels from auxiliary dataset. These approaches get the mask or pixels from Imagenet [1, 2], ADE 20k [12], and COCO [5, 41]. By adopting the regularization loss suggested by OE and using additional post-processing, Meta-OOD [5] enhances performance. PEBAL [41] also adopts the energy regularization loss proposed by EnergyOE and boost its performance through abstention learning and some additional regularization losses. DenseHybrid [12] utilizes a hybrid model of discriminative and generative classifiers. The following approach [1, 2] are based on a binary classifier.

2.3. OOD in Long-tailed Image Classification

Real-world data frequently exhibits a long-tail distribution, and learning from such imbalanced data has been questioned [13]. Deep neural networks demonstrate the degradation of performance when training on data with class imbalance [4]. There are primarily two approaches resolving the issue of class imbalance. The first is a technique for readjusting the weights for each sample in the training loss [4, 39]. The second approach is a method to train multiple expert models to ensemble [44].

PASCL [43] tackles the OOD detection problem and finds difficulty in the long-tailed training set. Similarly, [30] deals with the open set classification challenge in the long-tailed training set. Particularly, PASCL assesses the performance of MSP [17], OE [18], EnergyOE [29], SOFL [35], OECC [38], and NTOM [6] in the SC-OOD benchmark [46] as a baseline for the OOD detection problem in long-tailed image classification. By incorporating partiality and asymmetry to the existing supervised contrastive learning to accommodate the long-tailed situation, PASCL achieves SOTA performance.

3. Proposed Method

3.1. Preliminary

We can formalize a discriminative neural classifier as $f(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^K$,which maps an input image $\mathbf{x}$ with D dimension to a real-valued vector (logit) with K dimension which is a number of classes. Probability vector $F(\mathbf{x})$ is computed as $Softmax(f(\mathbf{x}))$, which satisfies $\mathbf{1}^T F(\mathbf{x}) = 1$ and $F(\mathbf{x}) \geq 0$. $y \in \{1, 2, \dots, K\}$ represents the class label. $f_y(\mathbf{x})$ and $F_y(\mathbf{x})$ indicates the y th index of $f(\mathbf{x})$ and $F(\mathbf{x})$, respectively .

Outlier Exposure(OE) [18] leverages the cross-entropy loss for the existing training (in-distribution) data and the regularization loss for the auxiliary (OOD) data. The minimization goal for the maximum softmax probability baseline detector is as follows:

$$\min_{\theta} \quad \mathbb{E}_{(\mathbf{x},y)\sim D_{in}^{train}}[-\log F_y(\mathbf{x})] + \lambda L_{OE}, \tag{1}$$

where $L_{OE} = \mathbb{E}_{\mathbf{x}_{out}\sim D_{out}^{train}}[H(\mathbf{u}; F(\mathbf{x}))]$, whereas D_{in} and D_{out} denote the in-distribution(ID) training set and the OOD training set, respectively. $\mathbf{u}$ is uniform distribution and H is cross entropy loss.L_{OE} is a regularization loss for OE

Energy-based OOD detection (EnergyOE) also leverages the cross-entropy loss for the training (ID) data and the regularization loss for the auxiliary (OOD) data. However, EnergyOE proposes energy regularization loss L_{energy} different from that of L_{OE}, which is given by

$$\begin{aligned} L_{energy} &= L_{in,hinge} + L_{out,hinge} \\ &= \mathbb{E}_{(\mathbf{x}_{in},y)\sim D_{in}^{train}}[(\max(0, E(\mathbf{x}) - m_{in}))^2] \\ &+ \mathbb{E}_{\mathbf{x}_{out}\sim D_{out}^{train}}[(\max(0, E(\mathbf{x}) - m_{out}))^2], \end{aligned} \tag{2}$$

where $E(\mathbf{x}; f) = -T \cdot \log(\sum_{j=1}^{K} e^{f_j(\mathbf{x}))/T})$. Energy function $E(\mathbf{x}; f)$ is computed as LogSumExp of logit with temperature scaling, In most cases, temperature T=1. Energy regularization loss is the sum of squared hinge losses for energy with each of the existing (ID) data and the auxiliary (OOD) data.

3.2. Balanced Energy regularization loss

Given the property of an OOD sample, our balanced energy regularization loss performs various regularizations for the OOD training data. The property of the OOD sample is modeled in the new term Z, which measures whether a sample belongs to a majority class or a minority class. The MAIN idea of our loss is to use a larger regularization to OOD samples of majority classes compared to OOD samples of minority classes.

For the OOD data, we created the Z term to measure whether a sample is of the majority or minority class. We require the prior probability of the OOD distribution to determine which class is the majority. Through inference on the pre-trained model of OOD data represented as auxiliary data, we obtain N_i, which is the number of samples that are classified as class i. Next, the prior probability of the OOD distribution is estimated by

$$P(y=i|o) = \frac{N_i}{N_1 + N_2 + \cdots + N_K}. \tag{3}$$

Using the discriminative neural classifier f, the posterior probability of the i-th class for a given image $\mathbf{x}$ is obtained by the softmax on the output of f, that is,

$$P(y=i|\mathbf{x}, o) = \frac{e^{f_i(\mathbf{x})}}{\sum_{j=1}^{K} e^{f_j(\mathbf{x})}}. \tag{4}$$

The higher the posterior probability of i-th class for $\mathbf{x}$, the higher the probability that $\mathbf{x}$ belongs to i-th class. And the higher the prior probability of i-th class, the higher the probability that i-th class is a majority class. Hence the higher the product of $P(y=i|o)$ and $P(y=i|\mathbf{x}, o)$, the higher the possibility that $\mathbf{x}$ belongs to a majority class i. From this result, a metric Z to measure a possibility that $\mathbf{x}$ belongs to majority classes, is defined by

$$Z = \sum_{j=1}^{K} P(y=j|\mathbf{x}, o) P(y=j|o). \tag{5}$$

In addition, we model additional generalized prior probability using hyperparameter γ. The degree of prior difference between classes is controlled by the hyperparameter γ. Finally, the generalized version Z_γ is defined by

$$Z_\gamma = \sum_{j=1}^{K} P(y=j|\mathbf{x}, o) P_\gamma(y=j|o), \tag{6}$$

where $P_\gamma(y=i|o) = L^1 norm\{P^\gamma(y=i|o)\}$. For numerical stability, we apply L1-normalization after multiplying prior probability $P(y=i|o)$ by itself γ times. If γ=0, then we model uniform prior probability and Z_γ becomes constant value $\frac{1}{K}$. If γ is negative, we model the inverse distribution of prior probability. As γ increases, the difference among prior probabilities of classes increases. Based on the Z_γ term, we design our balanced Energy regularization loss as follows.

$$\begin{aligned} L_{energy,bal} &= L_{in,hinge} + L_{out,bal} \\ &= \mathbb{E}_{(\mathbf{x}_{in},y)\sim D_{in}^{train}}[(\max(0, E(\mathbf{x}) - m_{in}))^2] \\ &+ \mathbb{E}_{\mathbf{x}\sim D_{out}^{train}}[(\max(0, E(\mathbf{x}) - m_{out} - \alpha Z_\gamma))^2 Z_\gamma], \end{aligned} \tag{7}$$

where $E(\mathbf{x}; f) = -T \cdot \log(\sum_{j=1}^{K} e^{f_j(\mathbf{x}))/T})$. Our loss $L_{energy,bal}$ is the sum of $L_{in,hinge}$ and $L_{out,bal}$. $L_{in,hinge}$ is squared hinge loss for in-distribution data, which is same as in Eq. (2). $L_{out,bal}$ is our novel loss with two adaptive loss components that depend on Z_γ. The margin is the first component and the weight is the second component. As Z_γ of a training sample increases, the loss margin and loss of weight increase, thus increasing the overall loss. First, the adaptive loss margin provides an additional Z_γ-proportional margin in the squared hinge loss for auxiliary data. As a result, our adaptive loss margin is defined by $\alpha \cdot Z_\gamma$ which is Z_γ multiplied by hyperparameter α. Second, the adaptive loss weight gives a weight proportional to Z_γ for the squared hinge loss. Thus, the squared hinge loss is multiplied by our adaptive loss weight Z_γ at the end.

3.3. Training Procedure

```
Algorithm 1: Balanced Energy Learning
Input: $f$:Pre-trained model
Data: $D_{in}$:in-distribution training set,
      $D_{out}$:OOD training set
Step1: Inference on OOD training set
Load the weight of pre-trained model $f$;
$N_j \longleftarrow 0$, for all $j$=1 to $K$
for $t = 1$ to $T_1$ do
    Sample a mini batch $D_{mini,o}$ from $D_{out}$
    Inference on the mini batch $f(D_{mini,o})$
    for $j = 1$ to $K$ do
        $n_j \longleftarrow$ count($\max_i f(D_{mini,o}), j$)
        $N_j \longleftarrow N_j + n_j$
Compute prior probability of OOD as Eq. (3).
Step2: Fine-tuning the pre-trained model
for $t = T_1 + 1$ to $T_2$ do
    Sample mini-batches $D_{mini,i}$ and $D_{mini,o}$
    from $D_{in}$ and $D_{out}$, respectively.
    Update unfrozen classification layers of $f$
    by minimizing Eq. (8).
```

Our method leverages the cross-entropy loss for the existing training (ID) data and the regularization loss for the auxiliary (OOD) data as Outlier Exposure (OE). Therefore, our minimizing objective is as follows:

$$\min_\theta \quad \mathbb{E}_{(\mathbf{x},y)\sim D_{in}^{train}}[-\log F_y(\mathbf{x})] + \lambda L_{energy,bal}. \tag{8}$$

Balanced energy regularization loss $L_{energy,bal}$ is defined in Section 3.2.

Next, we summarize our balanced energy learning process in Algorithm. 1. Our approach is predicated on the idea that we have a model that has already been trained following standard neural network training (ST). Therefore, the input is a pre-trained neural classifier f by ST process. In the image classification task, D_{in} is an original training image set, D_{out} is an unlabeled image set of auxiliary data. In the semantic segmentation task, D_{in} is an original training pixel set, D_{out} is a pixel set that is synthesized by a cut-pasted OOD mask from auxiliary data. Finally, our approach entails two steps. The first step is to determine N_i by concluding model f, after which the prior probability of OOD is calculated. The process of fine-tuning using our balanced energy regularization loss is the second step.

4. Experiments

4.1. Experiment Settings

4.1.1 Dataset

Semantic segmentation task: We use Cityscapes [8] dataset as ID data. We use the object mask of COCO [27] dataset as auxiliary data. For the OOD test data, we use Fishyscapes [3] dataset and Road Anomaly [28] dataset.

Long-tailed image classification task: We use two long-tailed image classification datasets CIFAR 10-LT [4] and CIFAR 100-LT [4] as ID data. Imbalance ratio ρ=100 following [43]. we utilized TinyImages 80M [42] dataset as auxiliary data. For the OOD test data, we use six datasets (CIFAR [21], Texture [7], SVHN [37], LSUN [47], Places365 [49], and TinyImagenet [23]) introduced in the SC-OOD benchmark [46].

Image classification task: We use CIFAR10 [21]and CIFAR100 [21] as ID data, and the rest are the same as in the case of long-tailed.

4.1.2 Model

Following [41], we use the semantic segmentation model of Deeplabv3+ like architecture with a WideResNet38 [50]. In long-tailed image classification and image classification task, we use ResNet18 [14] model as in [43]. To confirm generality in the long-tailed image classification task, We also use the WideResNet (WRN-40-2) [48] model.

4.1.3 Implementation Details

In semantic segmentation task, we employ a similar method as PEBAL [41]. we load the semantic segmentation pretrained model by NVIDIA [50] on the Cityscapes dataset. As in PEBAL, we build the auxiliary data by cutting and pasting the mask from the COCO data. The prior probability is then drived from OOD pixels for random sample of 10000 scene images. We use the same training configuration as PEBAL for fine-tuning, with 20 epochs, Adam as the optimizer, and a learning rate of 0.00001. The distinction is that batch size is configured to be 8.

In both long-tailed and normal image classification, we employ a similar method as EnergyOE [29]. By using the ST method, we can obtain a pre-trained model following the setting of OE [18]. Our auxiliary dataset is a subset of TinyImages80M with 300K images. Next, 300K images are used to extract the prior probability. We only use 30K subset images for training following PASCAL [43]. For fine-tuning, we use an almost identical training setting as EnergyOE, where the initial learning rate is 0.001 with cosine decay [32] and the batch size is 128 for in-distribution data and 256 for unlabeled OOD training data. The difference is that the number of epochs is 20 for a long-tailed task. We follow the hyperparameter settings of [29, 41]. We summarize our hyperparameter setting for all tasks in Table 1.

Table 1. Hyperparameter(hyper.P) setting in all tasks: Semantic segmentation (Seg), Long-tailed image classifcation(Long-tailed Cls), and image classification(Cls)

TASK	Seg	Long-tailed Cls		Cls	
hyper.P		CIF-10	CIF-100	CIF-10	CIF-100
class num K	19	10	100	10	100
temp T	1	1	1	1	1
λ	0.1	0.1	0.1	0.1	0.1
α	5	10	100	10	100
γ	3.0	0.75	0.75	0.025	0.025
m_{in}	-12	-23	-27	-23	-27
m_{out}	-6	-5	-5	-5	-5

Table 2. Evaluation result on Fishyscapes test sets (Lost&Found, Static) : OOD detection performance with AP and FPR. Compared Methods are: MSP; En† (Entropy); kNN† (kNN Embedding - density) SML; BD† (Bayesian Deeplab); DSN† (Density Single-layer NLL); DMN† (Density Minimum NLL); IR† (Image Resynthesis); DLR† (Density Logistic Regression); SB† (SynBoost); DODH† (Discriminative Outlier Detection Head); OTVC† (OoD Training - Void Class); DD† (Dirichlet Deeplab); DH† (DenseHybrid); PEBAL; **Ours: (Balanced Energy PEBAL)**

$R^\dagger$: Re-training, $E^\dagger$: Extra Network, $O^\dagger$: OoD Data.

Method	$R^\dagger$	$E^\dagger$	$O^\dagger$	FS Lost & Found		FS Static	
				AP↑	FPR↓	AP↑	FPR↓
MSP [16]	✗	✗	✗	1.77	44.85	12.88	39.83
En† [17]	✗	✗	✗	2.93	44.83	15.41	39.75
kNN† [3]	✗	✗	✗	3.55	30.02	44.03	20.25
SML [19]	✗	✗	✗	31.05	21.52	53.11	19.64
BD† [36]	✓	✗	✗	9.81	38.46	48.70	15.05
DSN† [3]	✗	✓	✗	3.01	32.90	40.86	21.29
DMN† [3]	✗	✓	✗	4.25	47.15	62.14	17.43
IR† [28]	✗	✓	✗	5.70	48.05	29.60	27.13
DLR† [3]	✗	✓	✓	4.65	24.36	57.16	13.39
SB† [10]	✗	✓	✓	43.22	15.79	72.59	18.75
DODH† [2]	✓	✓	✓	31.31	19.02	96.76	0.29
OTVC†	✓	✗	✓	10.29	22.11	45.00	19.40
DD† [33]	✓	✗	✓	34.28	47.43	31.30	84.60
DH† [12]	✓	✗	✓	47.06	3.97	80.23	5.95
PEBAL [41]	✓	✗	✓	44.17	7.58	92.38	1.73
Ours	✓	✗	✓	**51.83**	**3.76**	**94.62**	**0.99**

4.2. Semantic Segmentation

Table 2 shows the results of our approach on the Fishyscapes leaderboard. The technique that replaces the energy regularization loss in PEBAL with our balanced energy regularization loss is known as our balanced energy PEBAL. Our approach outperforms PEBAL and achieves SOTA in a methodology that utilize OOD data and require no extra network.

Table 3 presents the results of our method on the Fishyscapes validation sets and Road Anomaly test set.

Table 3. Evaluation result on Fishyscapes validation sets and Road Anomaly test set : OOD detection performance with AUROC, AP, and FPR

Method	FS Lost & Found			FS Static			Road Anomaly		
	AUC↑	AP↑	FPR↓	AUC↑	AP↑	FPR↓	AUC↑	AP↑	FPR↓
MSP [16]	89.29	4.59	40.59	92.36	19.09	23.99	67.53	15.72	71.38
Max Logit [16]	93.41	14.59	42.21	95.66	38.64	18.26	72.78	18.98	70.48
Entropy [17]	90.82	10.36	40.34	93.14	26.77	23.31	68.80	16.97	71.10
Energy [29]	93.72	16.05	41.78	95.90	41.68	17.78	73.35	19.54	70.17
Mahalanobis [24]	96.75	56.57	11.24	96.76	27.37	11.7	62.85	14.37	81.09
Meta-OOD [5]	93.06	41.31	37.69	97.56	72.91	13.57	-	-	-
Synboost [10]	96.21	60.58	31.02	95.87	66.44	25.59	81.91	38.21	64.75
SML [19]	94.97	22.74	33.49	97.25	66.72	12.14	75.16	17.52	70.70
Deep Gambler [31]	97.82	31.34	10.16	98.88	84.57	3.39	78.29	23.26	65.12
PEBAL [41]	98.96	58.81	4.76	**99.61**	92.08	1.52	87.63	**45.10**	44.58
Balanced Energy PEBAL (Ours)	**99.03**	**67.07**	**2.93**	99.55	**92.49**	**1.17**	**88.36**	43.58	**41.54**
EnergyOE [29]	98.14	45.61	8.21	99.32	89.12	2.62	83.32	32.59	53.01
Balanced EnergyOE (Ours)	**98.42**	**54.58**	**6.70**	**99.43**	**91.77**	**1.63**	**85.50**	**34.90**	**46.60**

Table 4. Evaluation result on CIFAR10-LT using ResNet18; (a): OOD detection performance with AUROC,AP and FPR; Mean over six random runs are reported(OE,EnergyOE,Ours). (b): Comparison result with other methods; average (over 6 datasets) OOD detection performance (AUROC,AP, FPR) and classification accuracy (ACC).

(a)

Dataset	Method	AUC↑	AP↑	FPR↓
Texture	OE (tune)	87.98	80.05	45.54
	EnergyOE (tune)	95.53	**92.93**	23.26
	Ours	**95.69**	92.38	**21.26**
SVHN	OE (tune)	92.10	95.52	27.37
	EnergyOE (tune)	96.63	98.46	14.52
	Ours	**97.74**	**98.89**	**9.87**
CIFAR100	OE (tune)	78.24	76.35	65.28
	EnergyOE (tune)	84.44	84.63	59.92
	Ours	**85.20**	**84.98**	**57.95**
Tiny ImageNet	OE (tune)	81.47	75.79	58.68
	EnergyOE (tune)	88.40	84.95	45.17
	Ours	**88.92**	**84.98**	**42.38**
LSUN	OE (tune)	86.19	85.85	54.49
	EnergyOE (tune)	94.00	**93.70**	26.96
	Ours	**94.48**	93.15	**23.88**
Places365	OE (tune)	84.27	93.84	59.08
	EnergyOE (tune)	92.51	97.14	32.88
	Ours	**93.35**	**97.23**	**28.25**
Average	OE (tune)	85.04	84.57	51.74
	EnergyOE (tune)	91.92	**91.97**	33.79
	Ours	**92.56**	91.94	**30.60**

(b)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	70.96	69.35	67.37	69.83
	Energy [29](ST)	75.93	72.91	61.00	69.83
	OECC [38]	87.28	86.29	45.24	60.16
	EnergyOE [29](scratch)	89.31	88.92	40.88	74.68
	OE [18](scratch)	89.77	87.25	34.65	73.84
	PASCL [43]	90.99	89.24	33.36	77.08
	Open-Sampling [45]	90.24	85.44	31.00	77.06
	OE [18](tune)	85.04	84.57	51.74	69.79
	EnergyOE [29](tune)	91.92	**91.97**	33.79	74.53
	Ours	**92.56**	91.94	**30.60**	76.22
	Ours+AdjLogit [34]	**92.56**	91.94	**30.60**	**81.37**

Table 5. Evaluation result on CIFAR100-LT using ResNet18; (a): OOD detection performance with AUROC,AP and FPR; Mean over six random runs are reported(OE,EnergyOE,Ours). (b): Comparison result with other methods; average (over 6 datasets) OOD detection performance (AUROC,AP, FPR) and classification accuracy (ACC).

(a)

Dataset	Method	AUC↑	AP↑	FPR↓
Texture	OE (tune)	66.29	51.98	84.04
	EnergyOE (tune)	79.56	70.88	68.60
	Ours	**82.10**	**73.09**	**64.19**
SVHN	OE (tune)	74.93	85.41	63.94
	EnergyOE (tune)	86.19	91.74	42.27
	Ours	**88.66**	**92.88**	**33.79**
CIFAR10	OE (tune)	59.44	56.34	84.70
	EnergyOE (tune)	**61.15**	**56.66**	**82.60**
	Ours	59.40	54.97	85.16
Tiny ImageNet	OE (tune)	66.24	51.07	80.04
	EnergyOE (tune)	70.78	55.90	74.43
	Ours	**71.42**	**56.52**	**74.22**
LSUN	OE (tune)	73.46	59.07	73.05
	EnergyOE (tune)	81.61	69.16	57.37
	Ours	**83.83**	**71.23**	**52.04**
Places365	OE (tune)	71.70	85.08	74.62
	EnergyOE (tune)	79.12	89.09	61.96
	Ours	**81.10**	**89.94**	**57.52**
Average	OE (tune)	68.68	64.83	76.73
	EnergyOE (tune)	76.40	72.24	64.54
	Ours	**77.75**	**73.10**	**61.15**

(b)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	60.26	57.58	84.00	38.74
	Energy [29](ST)	63.22	59.06	81.12	38.74
	OECC [38]	70.38	66.87	73.15	32.93
	EnergyOE [29](scratch)	71.10	67.23	71.78	39.05
	OE [18](scratch)	72.91	67.16	68.89	39.04
	PASCL [43]	73.32	67.18	67.44	43.10
	Open-Sampling [45]	74.46	69.49	66.82	39.86
	OE [18](tune)	68.68	64.83	76.73	38.93
	EnergyOE [29](tune)	76.40	72.24	64.54	40.65
	Ours	**77.75**	**73.10**	**61.15**	41.05
	Ours+AdjLogit [34]	**77.75**	**73.10**	**61.15**	**45.66**

Here, we compare our method with not only PEBAL, but also EnergyOE, which is a baseline that use only energy regularization loss. We show that our loss enhances both compared to using the original energy regularization loss. Furthermore, balanced energy PEBAL has superior performance compared to other baselines.

4.3. Long-Tailed Image Classification

Table 4a shows the CIFAR10-LT experiment results of comparison with the existing baselines (OE, EnergyOE) in detail for each data. We present the average of all over six random runs (OE, EnergyOE, Ours). 16 out of 18 show better performance than baseline in AUROC, AP, and FPR. Table 4b depicts the summary of the CIFAR10-LT experiment results which is the average performance over 6 datasets

Table 6. Evaluation result on CIFAR using ResNet18; average (over 6 datasets) OOD detection performance (AUROC,AP, FPR) and classification accuracy (ACC) (a): Result on CIFAR10 (b): Result on CIFAR100

(a)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	89.25	86.63	31.32	**93.69**
	Energy [29](ST)	91.55	89.88	29.07	**93.69**
	OECC [38]	96.33	95.38	**14.36**	91.57
	OE [18](tune)	95.68	95.36	18.20	93.37
	EnergyOE [29](tune)	96.77	**96.72**	14.82	93.30
	Ours	**96.83**	96.70	14.51	93.00

(b)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	76.14	71.29	62.78	**75.70**
	Energy [29](ST)	79.78	73.31	57.59	**75.70**
	OECC [38]	84.03	77.94	45.26	69.55
	OE [18](tune)	82.76	77.93	51.72	74.33
	EnergyOE [29](tune)	85.84	**80.99**	43.02	74.95
	Ours	**85.85**	80.91	**42.93**	74.83

compared to the other methods. When compared to the baseline results for the model before fine-tuning, our approach performs better. The point to note is that accuracy also improves compared to using the energy regularization loss. When evaluating the accuracy, PASCL has stage 2 which uses the loss proposed by AdjLogit [34] to improve long-tailed classification accuracy. For a fair comparison with PASCL, we also report our accuracy after going through stage 2 used by PASCL. Our method outperforms the SOTA algorithm, PASCL.

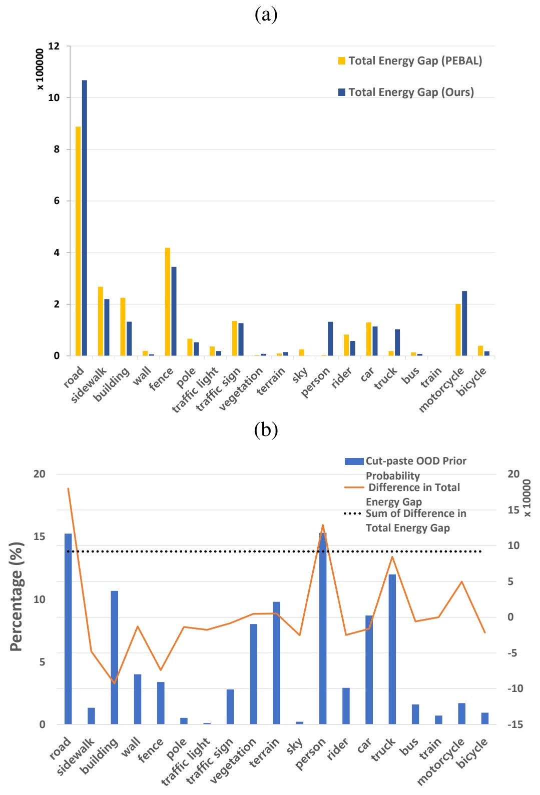
Table 5a details the comparison between the CIFAR100-LT experiment results and the current baselines (OE, EnergyOE). We present the average of all six random runs (OE, EnergyOE, Ours).15 out of 18 show better performance than baseline in AUROC, AP, and FPR. Table 5b presents the summary of CIFAR100-LT experiment results which is the average performance over 6 datasets compared to the other methods. Similarly, as CIFAR10-LT, accuracy also improves compared to using the energy regularization loss. Our method outperforms the SOTA algorithm, PASCL.

4.4. Image Classification

Similar to the long-tailed image classification task, we compare our approach with the existing baselines (OE, EnergyOE). Table 6a depicts the summary of CIFAR10 experiment results which is the average performance over 6 datasets compared to the other methods. The accuracy performs marginally worse than using energy regularization loss, which is different from long-tailed case. However, the OOD performance is still improved compared to using energy regularization loss. Table 6b illustrates the summary of CIFAR100 experiment results which are the average performance over 6 datasets compared to the other methods. Similar to the CIFAR10 experiment, OOD performance improves when using our losses compared to baseline.

5. Discussion

5.1. Empirical Analysis


(a)

(b)

Figure 2. Analysis results of our method in the semantic segmentation task (Fishyscapes Lost&Found validation set) (a): Comparison of the class-wise total energy gap of PEBAL and our method; (b): class-wise total energy gap difference between PEBAL and our method.

We define the novel term Energy Gap as $\mathbb{E}[E(x_{in})] - \mathbb{E}[E(x_{out})]$, which measures the average energy gap between ID data and OOD data. We can measure it class-wise and i th class Energy Gap is $\mathbb{E}[E(x_{in,i})] - \mathbb{E}[E(x_{out,i})]$. Finally, the class-wise Total Energy Gap is defined as $(\mathbb{E}[E(x_{in,i})] - \mathbb{E}[E(x_{out,i})]) \cdot N_{out,i}$ by multiplying frequency of class i OOD data. Intuitively, the larger the Total Energy Gap, the larger the energy gap between ID and OOD, and the better the OOD detection performance. Figure 2a compares the results of our method and PEBAL with regard to the class-wise Total Energy Gap. We see that this gap is elevated in the majority class, like the road.

Figure 2b shows the difference in Total Energy Gap between PEBAL and ours. As shown in Figure 2b, class-wise difference distribution shown in orange is similar to the prior probability shown in blue, which infers that our balanced energy regularization effectively works. Furthermore, by improving the gap on the majority class effectively, the sum of the class-wise Total Energy Gap increases (black dotted line is over 0). Thus, our method improves the OOD detection performance compared to PEBAL.

Table 7. Evaluation result in semantic segmentation task depending on γ : OOD detection performance(AUROC,AP,FPR) and accuracy(MIOU for Cityscapes validation) on the Fishyscapes validation sets

Method		City	FS Lost & Found			FS Static		
Name	γ	MIOU↑	AUC↑	AP↑	FPR↓	AUC↑	AP↑	FPR↓
EnergyOE	0.0	89.07	98.14	45.61	8.21	99.32	89.12	2.62
Balanced EnergyOE	1.0	89.04	98.01	50.42	8.87	99.34	89.28	2.33
	2.0	**89.83**	98.72	53.36	6.57	99.42	90.32	2.31
	3.0	88.91	98.42	**54.58**	6.70	**99.43**	**91.77**	**1.63**
	4.0	88.53	**98.81**	53.27	**5.18**	99.34	89.98	2.39
Inverse Balanced EnergyOE	-3.0	84.28	95.49	43.94	31.28	98.45	81.32	5.66

Table 8. Evaluation result on CIFAR10-LT depending on γ :average (over 6 datasets) OOD detection performance (AUROC,AP, FPR) and classification accuracy (ACC) with model ResNet18; Mean over six random runs are reported.

Method		ACC↑	Average (total 6)		
Name	γ		AUC↑	AP↑	FPR↓
Energy OE	0.00	74.53	91.92	91.97	33.79
Balanced Energy OE	0.10	75.03	92.01	91.57	32.80
	0.25	75.23	92.16	91.36	31.83
	0.50	75.92	92.44	91.67	30.81
	0.75	**76.22**	**92.56**	91.94	**30.60**
	1.00	74.85	92.45	**92.03**	31.86
	1.25	72.38	92.33	**92.03**	32.60
Inv-Balanced Energy OE	-0.75	64.24	90.75	90.83	39.82

5.2. Ablation Study

5.2.1 Hyperparameter Analysis

As discussed in Section 3.2, the hyperparameter γ controls the degree of prior difference between classes. We fix α as in Table 1 and find the best γ starting from base case of $\gamma = 0$. For accurate implementation of baseline EnergyOE, we only set $\alpha = 0$, when $\gamma = 0$. For semantic segmentation task and long-tailed image classification task as in Table 7 and Table 8, we obtain a common tendency for γ. First, as γ becomes larger than 0, the OOD detection performance and accuracy improves, then obtain an optimal value and decrease again. Second, in the inverse case where γ becomes smaller than 0, both OOD detection performance and accuracy are worse than the baseline. This would be evidence of the efficiency of prior probability.

5.2.2 Loss Component Analysis

Table 9. Results of the loss component ablation on Fishyscapes validation sets in semantic segmentation task

Loss Component		FS Lost & Found			FS Static		
Margin	Weight	AUC↑	AP↑	FPR↓	AUC↑	AP↑	FPR↓
✗	✗	98.14	45.61	8.21	99.32	89.12	2.62
✗	✓	98.31	47.10	7.04	99.34	89.73	2.31
✓	✗	**98.46**	51.65	7.17	99.41	90.14	2.23
✓	✓	98.42	**54.58**	**6.70**	**99.43**	**91.77**	**1.63**

As defined in Section 3.2, our balanced energy regularization loss has two adaptive loss components that depend on Z_γ. For the semantic segmentation task and long-tailed image classification task as in Table 9 and Table 10, ablation results on two loss components show that we can achieve the best performance when using both adaptive loss margin and adaptive loss weight.

Table 10. Results of the loss component ablation on both CIFAR10-LT and CIFAR100-LT in long-tailed image classification task

Loss Component		CIFAR10-LT Average			CIFAR100-LT Average		
Margin	Weight	AUC↑	AP↑	FPR↓	AUC↑	AP↑	FPR↓
✗	✗	91.92	91.97	33.79	76.40	72.24	64.54
✗	✓	92.32	91.39	31.00	77.28	72.20	61.92
✓	✗	92.32	**92.27**	32.44	77.17	72.99	63.23
✓	✓	**92.56**	91.94	**30.60**	**77.75**	**73.10**	**61.15**

5.2.3 Network Analysis

Table 11. Evaluation result on long-tailed CIFAR using WideResNet(WRN-40-2); average (over 6 datasets) OOD detection performance (AUROC,AP, FPR) and classification accuracy (ACC) (a): Result on CIFAR10-LT (b): Result on CIFAR100-LT

(a)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	74.59	72.62	63.96	72.32
	Energy [29](ST)	80.23	77.67	58.44	72.32
	OE [18](tune)	83.96	83.70	54.60	69.31
	EnergyOE [29](tune)	91.44	**91.01**	34.02	75.02
	Ours	**91.85**	90.48	**31.03**	**76.14**

(b)

Dataset	Method	AUC↑	AP↑	FPR↓	ACC↑
Average	MSP [17](ST)	60.24	57.22	83.52	40.96
	Energy [29](ST)	63.31	59.44	81.95	40.96
	OE [18](tune)	68.52	65.18	76.88	**41.73**
	EnergyOE [29](tune)	76.45	72.75	65.70	39.95
	Ours	**77.41**	**73.10**	**62.84**	39.44

To show the generality of our method, we perform a long-tailed image classification experiment on WideResNet (WRN-40-2) [48] instead of ResNet18 [14]. When γ is 0.5 (α as in Table 1), we get optimal performance in both CIFAR10-LT and CIFAR100-LT, and the results are as summarized in Table 11a and Table 11b, respectively. Our approach outperforms the current baselines (OE, EnergyOE) on both CIFAR10-LT and CIFAR100-LT, similar to the case with ResNet18.

6. Conclusion

To solve the OOD detection issue in various tasks, we propose a new balanced energy regularization loss. The main idea of our loss is to apply large regularization to auxiliary samples of majority classes, compared to those of minority. We show the effectiveness of our novel loss through extensive experiments on semantic segmentation, long-tailed image classification, and image classification datasets. **Limitations and potential negative social impacts** are provided in the supplement.

Acknowledgements

This work was supported by IITP/MSIT [No.B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis; No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], and KAIA/MOLIT [No.1615012984, Development of Automated Driving System (Lv.4/4+) based Car-Sharing Service Technologies].

References

[1] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018. 1, 3

[2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *German conference on pattern recognition*, pages 33–47. Springer, 2019. 1, 3, 5

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135, 2021. 1, 3, 5

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 3, 5

[5] 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, 3, 6

[6] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021. 3

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3, 5

[9] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 2

[10] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 1, 3, 5, 6

[11] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012. 1

[12] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. *arXiv preprint arXiv:2207.02606*, 2022. 1, 3, 5

[13] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 3

[14] 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. 5, 8

[15] 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. 1, 2

[16] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 1, 3, 5, 6

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 3, 5, 6, 7, 8

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 1, 2, 3, 5, 6, 7, 8

[19] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 1, 3, 5, 6

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 1

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[23] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[24] 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, 3, 6

[25] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017. 1

[26] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 1, 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[28] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019. 1, 3, 5

[29] 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, 3, 5, 6, 7, 8

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 1, 3

[31] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[33] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 1, 5

[34] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 6, 7

[35] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020. 3

[36] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 1, 3, 5

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[38] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021. 1, 2, 3, 6, 7

[39] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 735–744, 2021. 3

[40] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020. 1, 2

[41] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. *arXiv preprint arXiv:2111.12264*, 2021. 1, 2, 3, 5, 6

[42] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 5

[43] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022. 1, 2, 3, 5, 6

[44] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 3

[45] Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, pages 23615–23630. PMLR, 2022. 6

[46] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 1, 2, 3, 5

[47] 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. 5

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5, 8

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 5