Margin-bounded Confidence Scores for Out-of-Distribution Detection

Lakpa Tamang, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal
School of Information Technology, Faculty of Science, Engineering and Built Environment
Deakin University, Geelong, Victoria, Australia
l.tamang@research.deakin.edu.au, {reda.bouadjenek, richard.dazeley, sunil.aryal}@deakin.edu.au

***Abstract*—In many critical Machine Learning applications, such as autonomous driving and medical image diagnosis, the detection of out-of-distribution (OOD) samples is as crucial as accurately classifying in-distribution (ID) inputs. Recently *Outlier Exposure* (OE) based methods have shown promising results in detecting OOD inputs via model fine-tuning with auxiliary outlier data. However, most of the previous OE-based approaches emphasize more on synthesizing extra outlier samples or introducing regularization to diversify OOD sample space, which is rather unquantifiable in practice. In this work, we propose a novel and straightforward method called Margin bounded Confidence Scores (MaCS) to address the nontrivial OOD detection problem by enlarging the disparity between ID and OOD scores, which in turn makes the decision boundary more compact facilitating effective segregation with a simple threshold. Specifically, we augment the learning objective of an OE regularized classifier with a supplementary constraint, which penalizes high confidence scores for OOD inputs compared to that of ID and significantly enhances the OOD detection performance while maintaining the ID classification accuracy. Extensive experiments on various benchmark datasets for image classification tasks demonstrate the effectiveness of the proposed method by significantly outperforming state-of-the-art (S.O.T.A) methods on various benchmarking metrics. The code is publicly available at https://github.com/lakpa-tamang9/margin_ood**

***Index Terms*—Out-of-distribution, outlier exposure, confidence score, weighted penalty**

I. Introduction

Machine learning-based systems in critical applications such as autonomous driving and medical image diagnosis should equally prioritize accurately classifying in-distribution (ID) inputs and detecting out-of-distribution (OOD) samples, which are also referred to as anomalies or novelties. This issue may arise because real-world data are dynamic in nature, where distribution shifts frequently occur owing to the emergence of new classes, leading to significant differences in the posterior probabilities of input and labels [1]. Hence, a classification system must avoid classifying objects from unknown classes to establish user trust.

OOD detection is a classic yet essential ML problem that aims to resolve the fundamental issue of models being over-confident in classifying samples from different semantic distributions [5]. Hence, numerous approaches have been proposed to solve this task [6]–[11], which typically rely on a post-hoc detection strategy, employing thresholds or other criteria to identify OOD samples. Another technique that has attracted considerable attention is the Outlier Exposure (OE) method [3] that advocates the use of outliers to regularize the model and generate low confidence scores on unseen distributions. To compare the confidence scores, i.e., the maximum values of the Softmax probabilities of ID and OOD samples for some of these techniques, we refer to Fig. 1. Here, we train three image classification models – Maximum Softmax Probability (MSP) [2], OE [3], and our proposed method MaCS – on the CIFAR-100 dataset. We use test images from CIFAR-100 as ID samples and images from the iSUN dataset [4] as OOD samples. In the literature, these two datasets serve as popular benchmark datasets utilized for the OOD detection task; the former is primarily employed as ID, while the latter is used to represent OOD data. We employ boxplots for visualization and score comparison, from which we observe the following: First, the MSP method, a straightforward classification model that optimizes cross-entropy, exhibits overconfidence when applied to OOD samples as their scores overlap significantly with those of ID samples. Second, while OE generally helps to decrease the scores of OOD samples, the overlap between the scores of OOD and ID samples is still noteworthy. The reason for this is that outliers can occasionally produce confidence scores comparable to, or even higher than those of ID samples. As a result, OOD samples that lie in the decision boundary can be often falsely categorized as ID data, which poses a challenge in their clear separation.

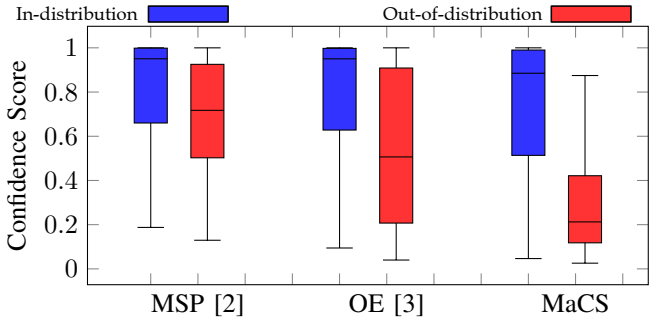

Fig. 1: Confidence scores of models trained using CIFAR-100 on test data from CIFAR-100 (ID samples) and iSUN [4] (OOD samples).

Moreover, most OOD detection methods rely on sampling and synthesizing existing outliers [12], [13], introducing regularization through augmentations [14], [15], and feature space maneuvering [16]. While these approaches attain reasonable

detection performance, they may often suffer from a phenomenon, which we refer to as "score explosion", where the confidence score for OOD samples exceeds that of ID samples as shown in Fig. 1. To address this issue, this paper introduces a novel approach called Margin bounded Confidence Scores (MaCS). Leveraging the insight gained from score explosions, MaCS penalizes the model during training, encouraging it to learn discriminative features between ID data and representative outliers. By nullifying score explosions and assigning weights based on the margin difference between ID and OOD confidence scores, MaCS aims to reduce model uncertainty in distinguishing between the two. In Fig. 1, the last two boxplots illustrate the distribution of scores for OOD and ID samples under MaCS, where clearly OOD samples receive significantly lower confidence scores compared to ID samples.

The **contributions** of this paper can be summarised as:

- **Simple and Practical Solution:** We investigate an OOD detection problem under a practical research setting, utilizing the existing confidence scores of any OE regularized model: a completely different approach compared to conventional outlier synthesis techniques whose objective is establishing heterogeneity of OOD sample space that cannot be quantitatively measured in practice.
- **Learning in Synergy:** We propose a novel and straightforward method called **Mar**gin bounded **C**onfidence **S**cores (MaCS) that work together with OE under a unified algorithm: a supplementary constraint is put forward to the training objective of the OE method to enhance the OOD detection robustness of a classification model.
- **Effectiveness:** We conduct comprehensive experiments utilizing established benchmark ID and OOD image classification datasets. Our findings reveal significant enhancements over several state-of-the-art (S.O.T.A) methods across various detection metrics. Furthermore, we validate our method by performing several ablation studies and prove it to be highly effective in achieving reliable detection performance under different networks, and datasets.

II. Related Works

There is a substantial body of research related to OOD detection techniques. Below, we review the major works related to post-hoc OOD detection and OOD detection by using auxiliary outliers.

Post-hoc OOD Detection: Post-hoc OOD detection techniques have the advantage of being easy to use without modifying the training procedure and objective of the model. In this regard, various scoring functions have been proposed to better utilize the high level semantic information of penultimate layers. A MaxLogit technique [17] uses the maximum value of logits instead of softmax probabilities to enhance the detection performance. In the following works, [18] used standarized value of maximum logit scores to align different distributions, and [19] decoupled the maximum logits value for flexibility to balance MaxCosine and MaxNorm. Similarly, ODIN [20] and Generalized ODIN [21] proposed the decomposition of confidence scores and modified input pre-processing methods to enhance detection performance. Additionally, ReAct [22] used activation rectification during the test time for stronger separation of ID and OOD data and DICE [23] used weight ranking to select the most salient weights to derive the OOD detection output.

OOD Detection by Using Auxiliary Datasets: Generating outliers or auxiliary OOD examples is essential to improve the robustness and generalization capabilities of a model [12]. The goal is to expose the model to a wider range of data scenarios beyond what is available in the training set. In literature, OOD detection has been realized by producing synthetic outliers using methods such as data augmentation [14], [24], [25], and adversarial example generation [26]–[29]. One such method, Energy OOD [30], uses energy scores instead of softmax scores because they are more aligned with the probability density of the inputs and are less prone to overconfidence. Another related study, GEM [31], models the feature space as a class-conditional multivariate Gaussian distribution. MixOE [15] and MiM [32] used MixUp regularizers to mix ID data with auxiliary outliers, with the former being in complex fine-grained scenarios. Motivated by the recent achievements of auxiliary outliers based approaches, our objective is to harness it's potential for OOD detection. Unlike other methods that depend partially or entirely on data augmentation-based regularization [15], [32] and intricate outlier synthesis/sampling techniques [12], [16], we present a less sophisticated method that relies on the confidence scores of a model while using eminent outlier datasets.

III. Background

A. Notation and Problem Definition

We consider a training dataset independently and identically distributed (*i.i.d*) data drawn from ID, $\mathcal{D}_{in} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(k)}, y^{(k)})\}$ with k instances, where each $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^n$ is an n-dimensional input feature vector of the instance i, and $y^{(i)} \in \mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$ represents its class. Similarly, during test phase, we evaluate the OOD detection capability using examples drawn from the OOD sample space $\mathcal{D}_{out} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(k)}\}$. Also, following the convention in [3], we introduce auxiliary outlier data as $\mathcal{D}_{out}^{OE}$ such that $\mathcal{D}_{out}^{OE} \cap \mathcal{D}_{out} \cap \mathcal{D}_{in} = \phi$.

The goal is then to learn a mapping function $f : \mathcal{X} \to \mathbb{R}^c$ trained using $\mathcal{D}_{in} \cup \mathcal{D}_{out}^{OE}$, which assigns to each feature vector $\mathbf{x}^{(i)} \in \mathcal{D}_{in}$ its correct class $y^{(i)}$, while avoiding classifying instances $\mathbf{x}^{(i)} \in \mathcal{D}_{out}^{OE}$.

B. Outlier Exposure

Outlier Exposure (OE), an auxiliary outlier based OOD detection method [3] is the baseline that we refer to in our study. It is a regularization technique that involves learning from additional datasets containing outliers or OOD samples with low confidence predictions along with standard training data. The goal is to expose the network to diverse OOD

examples during training, so that the model learns a more conservative concept of the ID data to distinguish them from their OOD counterparts. To achieve this, OE uses an auxiliary dataset of outliers $\mathcal{D}_{out}^{OE}$ that is entirely separate from the OOD test data $\mathcal{D}_{out}$. Hence, OE is trained by optimizing the following objective:

$$\mathcal{L}_{OE} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{in}}\left[\mathcal{L}(f(\mathbf{x}),y)\right] + \lambda_1 \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{out}^{OE}}[\mathcal{L}(f(\mathbf{x}),\mathcal{U})] \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss, $\mathcal{U} \in \mathbb{R}^k$ represents a uniform distribution over c classes, and λ_1 is the hyper-parameter for balancing both objectives.

C. Scoring Function

We adopt MSP as a method for detecting OOD samples, which operates using a threshold. MSP retains the maximum posterior probability (or confidence scores) over softmax probabilities of a network [2]. Thus, if $\mathbf{s}(\mathbf{x}) = \{s_1, s_2, ..., s_c\}$ denotes the confidence scores across c classes, the MSP is represented by $\max(\mathbf{s}(\mathbf{x}))$. In essence, by comparing this value with a predetermined threshold $\tau \in \{0, 1\}$, we can classify a given test input as either ID or OOD.

$$g(\mathbf{x}) = \begin{cases} \mathcal{ID}, & \text{if } \max(\mathbf{s}(\mathbf{x})) \geq \tau. \\ \mathcal{OOD}, & \text{otherwise.} \end{cases} \tag{2}$$

IV. Proposed Method: Margin bounded Confidence Scores (MaCS)

In this section, we introduce the MaCS framework. Initially, we augment Equation (1) with an additional loss component aimed at promoting a distinct separation between ID and OOD samples. Fig. 2 illustrates our approach, wherein we compute $\max(\mathbf{s}(\mathbf{x}))$ for inputs from both $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$, followed by subtracting the former from the latter. We refer to this operation as Maximum Confidence Difference (MCD), which is elaborated on in Section IV-A. Subsequently, we address score explosions, where the confidence score of the outlier exceeds that of the ID input. Finally, we constrain these score differences within a specified margin value. Further details regarding margin-based weighting are provided in Section IV-B.

A. Maximum Confidence Difference (MCD) and Penalty

We consider an input to the model, with equiproportionate samples from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$ such that a batch $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^{2N}$ has N samples from $\mathcal{D}_{in}$ and N samples from $\mathcal{D}_{out}^{OE}$ with the batch size of $2N$. We obtain confidence scores for $\mathcal{B}$ denoted as $\mathbf{S}_{\mathcal{B}} = \{\mathbf{s}(\mathbf{x}_i)\}_{i=1}^{2N}$. Next, we compute the maximum confidence score for each instance $\mathbf{x}_i \in \mathcal{B}$ as $\max(\mathbf{s}(\mathbf{x}_i))$. We denote these maximum scores as ID^{max} and OOD^{max} for inputs from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$, respectively. Note that both ID^{max} and OOD^{max} are N-dimensional vectors. Intuitively, the $\max(\mathbf{s}(\mathbf{x}_i))$ represents the notion of confidence of the model to categorize $\mathbf{x}_i$ into one of c classes. Subsequently, for each element of ID^{max} we compute the difference between every element of OOD^{max}. For instance, if there (see Fig. 2 for graphical illustration). We do this to ensure that every OOD input whose $\max(\mathbf{s}(\mathbf{x}_i))$ is larger than that of the ID is captured. Following that, we employ ReLU to penalize these occurrences by setting the negatives to zero, while retaining only the positives. Finally, the Maximum Confidence Difference (MCD) of batch $\mathcal{B}$ is estimated as:

$$\mathcal{MCD}(\mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\max(0, ID_i^{max} - OOD_j^{max})^2 \tag{3}$$

B. Bounded Margin

Furthermore, we bound the overall MCD term to be within a specified range to distinctly dispel the ID and OOD data thus subtracting it from the margin value. The idea is that for a correctly-classified ID sample, the model should not only be confident about it being correctly classified but also confident that it is not an OOD. Thus, we aim to give considerable attention to the OOD samples by assigning a weight $\mathcal{W}_{MaCS}$. We follow a similar idea of the weighting approach [33], [34], which attempts to solve class imbalance problem in classification tasks. Similar to how weights are administered to make the model more sensitive towards under-sampled classes, we attempt to assign weights to rectify the exploded scores. In particular, we typically assign weights based on the occurrence of score explosions, instead of class memberships [35], [36]. This phenomenon is crucial for OOD detection, where failing to detect an OOD sample is considered as severe as misclassifying an ID sample. With this, we define a more tailored weighting strategy that explicitly addresses the nature of the error, which is OOD scores exceeding ID scores rather than focusing on the under-represented classes. Mathematically, we administer $\mathcal{W}_{MaCS}$ as follows:

$$\mathcal{W}_{MaCS} = \max(0, m - \mathcal{MCD}(\mathcal{B})) \tag{4}$$

where m is the margin that enforces the minimum difference between the ID and OOD output. The best value of m is determined empirically and as explained in Section VII-A. To put this into perspective, if MCD goes to zero, we replace it with $\mathcal{W}_{MaCS}$, which relates to weight assignment for score explosions. As a supplement to the training objective of OE, we combine the term in (4) with (1) resulting in our final training objective for the whole dataset with B batches as follows:

$$\mathcal{L}_{final} = \sum_{i=1}^{B}(\mathcal{L}_{OE}^{(i)} + \lambda_2 \mathcal{W}_{MaCS}^{(i)}) \tag{5}$$

where, λ_2 is a hyperparameter for balancing the effect of weighted margin on $\mathcal{L}_{OE}$. We summarize the whole procedure of fine-tuning MaCS as a pseudo-code in Algorithm 1.

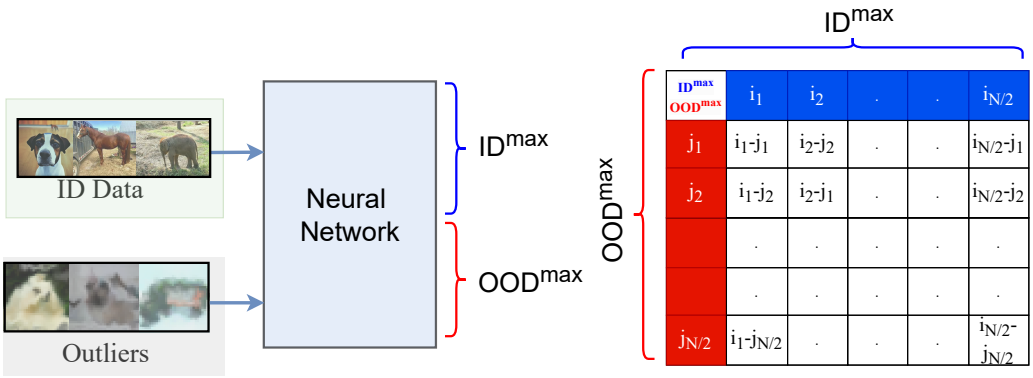

Fig. 2: Schematic overview of MaCS where the maximum confidence scores of inputs from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$ are extracted from the output layer of neural network followed by element-wise difference computation between ID^{max} and OOD^{max}.

Algorithm 1 Fine-tuning Margin Bounded Confidence Scores

Input: $\mathcal{D}_{in}$, $\mathcal{D}_{out}^{OE}$, pre-trained model f, hyperparameters θ, epochs T, and margin m;
Output: finetuned model f^* with θ^*, and m^*;

for m = 0.0 to 0.9 with step-size of 0.1 **do**
  for epoch = 1 to T **do**
    for batch = 1 to B **do**
      Select a batch $\mathcal{B} = 2N$, with N outliers, and N ID inputs from $\mathcal{D}_{out}^{OE}$, and $\mathcal{D}_{in}$ respectively;
      Concatenate sampled data from $\mathcal{D}_{in}$, and $\mathcal{D}_{out}^{OE}$ to create new input data, $\mathbf{x}$;
      Calculate $f(\mathbf{x};\theta)$, to get confidence scores;
      Compute maximum confidence score for each input with MSP;
      Compute MCD using (3);
      Compute $\mathcal{W}_{MACS}$ using (4)
      Compute overall loss using (5)
    end for
  end for
end for

V. Experiments and Results

This section outlines our experimental setup for conducting methodological evaluation, which include details regarding the benchmark datasets, baselines, and metrics utilized in our analysis.

A. Datasets

We categorize our data into three types: ID, outlier, and OOD datasets. The ID and outlier datasets are explicitly designated for training or fine-tuning purposes, while the OOD datasets are reserved for testing scenarios only.

1) ID Datasets

Our experiments are performed on four different image datasets: (1) **CIFAR-10** [37]: A small image classification dataset with 10 classes; (2) **CIFAR-100** [37]: A medium-scale image classification dataset with 100 classes; (3) **SVHN** [38]: A small-scale image dataset with 10 classes, consisting of digits from 0 to 9; and (4) **Imagenet-32** [39]: A down-sampled version of the original Imagenet-1k [40], which is considered a large-scale dataset due to its 1,000 classes. Note that our training, validation, and test data follow the standard splits provided.

2) Outlier Datasets

As an outlier dataset, earlier works have adopted 80 Million Tiny Images [41]; however, it has recently been advised by [42] that due to the presence of biases, offensive and prejudicial images it's further usage has been discontinued. Considering the ethical research practice, we therefore, use 300K Random Images, which is a de-biased subset of the same prepared by [3].

3) OOD Datasets

We follow the baseline works [3], [30] to adapt the common OOD dataset benchmarks. These include Textures [43], LSUN-C [44], SVHN [38], iSUN [4], and Places365 [45]. We only use the test sets of these data as OOD datasets.

B. Baseline and S.O.T.A Approaches

We compare our method with four different competitive baseline OOD detection approaches: (1) **OE** [3], and remaining three are it's variants that follow the similar principle of model regularization by training with auxiliary outliers; (2) **Energy** [30] (2020) employs energy scores aligned with the probability density of inputs for OOD detection; (3) **MixOE** [15] (2023) utilizes Mixup technique to mix $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$ to further enhance model regularization; and (4) **DivOE** [13] (2024) diversifies $\mathcal{D}_{out}^{OE}$ by explicitly synthesizing more informative outliers for extrapolation during training. We re-implemented these methods using their publicly available source codes, following the datasets and training configurations described in Sections V-A, and V-C respectively.

C. Training Configuration

In general, OE and its variants are trained in a fine-tuned scenarios. This approach is more practical because it is more

common to equip deployed models with the ability to detect OOD inputs rather than training a dual task (ID classification and OOD detection) from scratch. Following a similar setup, we use pre-trained baselines for models that are available. For models that do not have a pre-trained baseline, we initially train the model from scratch using a MSP [2] objective and then utilize it for fine-tuning.

Models and Hyperparameters: We train our method on four different neural network (backbone) architectures that are considered pre-eminent in image classification tasks; WideResnet [46], Allconv [47], Resnet [48], and Densenet [49]. For the sake of equivalence comparison with OE [3], we use their default hyperparameters. Specifically, for WideResnet architecture we use a total of 40 layers with a widen factor of 2, and dropout rate of 0.3. Likewise, we use Allconv with 9 layers, each comprising a combination of (Conv2D - BatchNorm2D - GELU). Furthermore, we use Resnet and Densenet models with 18 and 121 layer variants respectively. All the networks are fine-tuned on a pre-trained model upto 10 epochs using a stochastic gradient descent (SGD) optimizer with weight decay of $5e-4$, an initial learning rate of 0.001 with cosine decay. Unlike [3] that employed varying sample sizes for $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{OE}$, our approach utilizes equivalent sample sizes of $N = 128$, with a cumulative batch size of $\mathcal{B} = 2N = 256$ to enable post-hoc calculations.. The choices of λ_1, and λ_2 are both set at 0.5. Lastly, we select the value of $m \in \{0.1, 0.2, \ldots, 0.9\}$. All experiments were conducted on multiple RTX A4000 GPU servers.

D. Evaluation Metrics

We evaluated the detection performance using several metrics: (i) **AUROC:** It measures the discriminative capability of an OOD classifier in discerning ID and OOD data. It's value ranges from 0 to 1, the latter indicating perfect distinction. (ii) **AUPR:** This metric evaluates the trade-off between precision and recall, usually under class imbalance scenarios. Higher value of AUPR indicates better detection performance. (iii) **FPR95:** This metric is significant for assessing the robustness of OOD detection under high recall conditions. Ideally, a lower value of FPR95 is desirable which indicates fewer ID samples are incorrectly classified as OOD. We also evaluate the classification performance of the ID inputs using accuracy metric represented as ID-ACC.

VI. Results Comparison

In this section, we compare our results with the baseline and S.O.T.A methods as discussed in Section V-B. Across all metrics, we report an averaged performance and a standard error value that was determined through the execution of 10 independent test trials.

A. Detection Results

First, we present the detection results. Here, we test on the fine-tuned methods on same backbone of Wideresnet architecture with specifications as stated in Section V-C.

Table Ia presents the detection metrics with CIFAR-10 and CIFAR-100 as ID datasets. We present results of our method in two variants (same across all experiments, hereafter): MaCS and MaCS*, the former one fine-tuned at fixed m = 0.5, while the latter fine-tuned with respective optimal value of m for each test setting as reported in Table. III. From the results in Table. Ia, we can observe that MaCS* consistently outperformed all other methods across both ID datasets, not only in terms of detection metrics but also proving effective in generously classifying ID samples. MaCS was also able to obtain good detection performance coming second to MaCS* with CIFAR-100 ID data. The key reason for the improved performance can be attributed to the weighted penalization feature of MaCS. Because the model is trained to focus entirely on the score explosions, it becomes apparent that the model learns to restrict OOD scores to be smaller than that of ID scores. The comprehensive results on CIFAR ID benchmarks for each test OOD dataset evaluated under different methods with different backbone architectures are listed in Table IIa.

Similarly, in Table Ib, we compare our results by changing the ID inputs from CIFAR datasets to SVHN and Imagenet-32 but fine-tuned on the same architecture. Analyzing the results, it is evident that MaCS* performance remains superior regardless of change in $\mathcal{D}_{in}$. For SVHN, MaCS* reports FPR95 value to be as low as zero, while for large-scale Imagenet-32 we beat OE, and Energy [30], the second best method by 4.58 %, and 0.64% respectively. Note that for tests conducted with SVHN and Imagenet-32 as ID datasets, the results represent average scores across all OOD datasets except SVHN. Considering only the confidence score based supplementary constraint to conventional OE's objective, the gain in OOD detection performance is substantial. Interestingly, it is noteworthy to realize that MaCS and MaCS* were able to outperform relatively sophisticated methods such as MixOE [15] and DivOE [13]. This demonstrates the effectiveness of the proposed method which in addition to being conceptually simpler also yields exquisite performance. The comprehensive results on SVHN and Imagenet-32 ID benchmarks for each test OOD dataset evaluated under different methods with different backbone architecture are listed in Table IIb.

On other note, while training to distinguish ID and OOD samples based on their confidence scores, our method simultaneously learns to make the inter-class decision boundary of ID samples more compact, leading to fewer classification errors. The rationale behind this is that, with the cost function being penalized for every score explosion, the model takes wise decision in mapping inputs to corresponding distributions while keeping the loss value down throughout.

B. Confidence Scores Disparity between ID and OOD Data

MaCS's objective is to penalize score explosions, with the aim of increasing the disparity between ID and OOD scores. This is intended to make the separation between the two more apparent when thresholding with (2). To illustrate this property of MaCS, we trained two different backbone architectures, WRN and Allconv, using CIFAR-100 as ID data.

TABLE I: Comparison of OOD detection results on different ID datasets fine-tuned on a WRN architecture using 300K Random Images as auxiliary outliers. Best and second best values are reported in bold, and underline respectively. Arrows represent the direction towards optimum value.

Method \ ID Data	CIFAR-10				CIFAR-100			
	AUROC ↑	**AUPR ↑**	**FPR95 ↓**	**ID-ACC ↑**	**AUROC ↑**	**AUPR ↑**	**FPR95 ↓**	**ID-ACC ↑**
OE [3]	98.65±0.03	98.6±0.05	6.21±0.13	94.83±0.06	88.51±0.15	87.43±0.16	42.12±0.44	75.75±0.11
Energy [30]	98.68±0.03	98.49±0.05	5.88±0.13	94.35±0.07	87.567±0.06	87.77±0.09	48.93±0.19	74.77±0.11
MixOE [15]	90.85±0.12	90.48±0.2	41.46±0.36	94.53±0.03	78.02±0.22	73.98±0.29	61.34±0.38	75.17±0.18
DivOE [13]	98.46±0.04	98.38±0.05	7.15±0.19	95.01±0.05	87.42±0.08	86.45±0.06	44.21±0.27	75.83±0.09
MaCS	98.79±0.02	98.77±0.03	5.14±0.11	95.28±0.06	89.43±0.08	88.82±0.15	41.52±0.29	75.53±0.07
MaCS*	**98.79±0.02**	**98.77±0.03**	**5.14±0.11**	**95.28±0.06**	**90.93±0.13**	**90.28±0.21**	**37.54±0.35**	**76.12±0.04**

(a) CIFAR-10 and CIFAR-100

Method \ ID Data	SVHN				Imagenet-32			
	AUROC ↑	**AUPR ↑**	**FPR95 ↓**	**ID-ACC ↑**	**AUROC ↑**	**AUPR ↑**	**FPR95 ↓**	**ID-ACC ↑**
OE [3]	99.93±0.0	99.94±0.0	0.11±0.01	94.67±0.04	88.76±0.02	87.21±0.02	39.72±0.09	34.42±0.06
Energy [30]	99.92±0.0	98.93±0.0	0.14±0.01	94.35±0.03	90.88±0.02	89.53±0.03	31.3±0.1	32.26±0.06
MixOE [15]	96.98±0.04	96.44±0.08	13.21±0.1	88.59±0.36	72.56±0.09	64.0±0.09	59.83±0.11	31.66±0.05
DivOE [13]	99.95±0.0	99.95±0.0	0.04±0.0	94.66±0.02	90.44±0.02	89.14±0.03	35.52±0.05	34.34±0.05
MaCS	99.97±0.0	99.97±0.0	0.03±0.0	95.2±0.03	91.49±0.03	90.47±0.03	30.66±0.09	38.11±0.1
MaCS*	**99.98±0.0**	**99.98±0.0**	**0.0±0.0**	**95.4±0.02**	**91.49±0.03**	**90.47±0.03**	**30.66±0.09**	**38.11±0.1**

(b) SVHN and Imagenet-32

We then plotted the Kernel Density Estimation (KDE) plot of the confidence scores for two OOD test datasets: SVHN and iSUN, as shown in Fig. 3. As can be seen from the figure, the confidence scores for ID data are higher and close to 1, while those for OOD data are close to 0. Interestingly, we can also see that the overlap between these scores for MaCS is lower than that of OE, indicating that MaCS is better at distinguishing between ID and OOD samples. Given that MaCS penalizes score explosions and limits them to a defined margin, (i) OOD scores tend to be lower than their ID counterpart, and (ii) a sufficient gap (equivalent to m) between ID and OOD scores is ensured.

VII. Ablation Study

In this section we describe multiple experiments performed to evaluate the contributions made by the individual components of the proposed method.

A. Effect of Margin on the Detection Performance

In this ablation study, we evaluated the detection performance of the proposed method under different values of m. We used a margin value $m \in \{0.0, 0.1, \ldots, 0.9\}$, and fine-tuned two models WRN and Allconv using all four ID datasets as mentioned in Section. V-A1. Fig. 4, depicts the AUROC, AUPR, and FPR95 scores averaged over five different test OOD datasets against the range of values of m. From the figure, we can observe that the characteristics of the curve remains different for different ID datasets, nonetheless, for a particular ID dataset both models (WRN, and Allconv) exhibit similar trend throughout the values of m. Overall, the model is seen to perform best at or after $m = 0.5$. In terms of the impact of m, most of the time larger values are expected to increase the dispersion of OOD and ID scores towards their respective likelihood limits of 0 and 1. We record the optimum detection results for each dataset, across both models and report it in Table. III. These results emphasize the importance of carefully selecting the value of m to achieve optimal performance for MaCS.

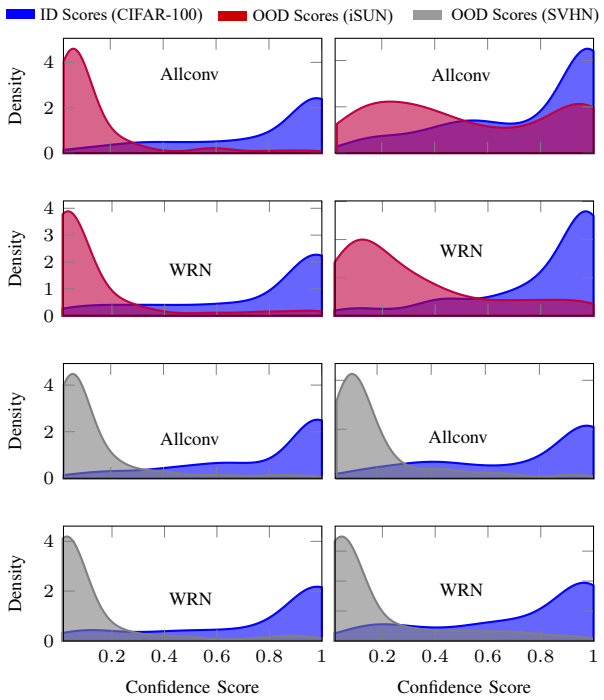

Fig. 3: KDE plot of confidence scores for two OOD test data: iSUN and SVHN against CIFAR-100 ID data trained on a WRN architecture. Left column plots are for MaCS, and right column plots are for OE.

TABLE II: Comprehensive OOD detection results comparison of MaCS on different ID datasets with S.O.T.A methods. All methods are trained on a WRN architecture. Best, and second best results are represented in bold and underline respectively.

		CIFAR-10						CIFAR-100					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
Textures	OE [3]	98.40±0.05	98.35±0.06	8.49±0.28	97.67±0.04	97.73±0.04	13.21±0.30	86.22±0.18	85.53±0.21	49.84±0.64	80.82±0.14	79.22±0.22	60.12±0.33
	Energy [30]	98.68±0.03	98.56±0.04	6.41±0.28	97.27±0.04	97.44±0.04	14.32±0.31	84.39±0.07	85.05±0.08	59.49±0.45	77.67±0.19	77.59±0.20	68.19±0.85
	MixOE [15]	87.91±0.12	87.87±0.20	60.21±0.70	93.04±0.07	92.19±0.07	26.58±0.32	76.15±0.24	72.33±0.32	68.73±0.50	74.95±0.16	71.17±0.19	71.33±0.61
	DivOE [13]	98.25±0.05	98.23±0.05	9.37±0.31	97.63±0.05	97.74±0.06	13.78±0.34	87.08±0.11	86.97±0.1	48.37±0.36	81.34±0.19	81.02±0.2	60.05±0.48
	MaCS	98.74±0.03	98.76±0.03	5.07±0.16	98.29±0.04	98.28±0.04	9.60±0.23	88.15±0.11	88.14±0.12	46.89±0.43	81.84±0.16	80.42±0.25	58.73±0.57
	MaCS*	**98.74±0.03**	**98.76±0.03**	**5.07±0.16**	**98.69±0.02**	**98.72±0.03**	**7.68±0.18**	**88.61±0.18**	**88.49±0.23**	**46.00±0.60**	**83.81±0.20**	**83.31±0.14**	**57.18±0.69**
iSUN	OE [3]	99.05±0.04	98.88±0.07	4.60±0.16	98.17±0.04	98.03±0.04	9.30±0.25	84.79±0.14	83.38±0.14	52.81±0.49	68.81±0.2	70.01±0.22	82.22±0.39
	Energy [30]	99.10±0.03	98.83±0.05	3.38±0.14	96.50±0.06	96.24±0.07	15.54±0.32	86.95±0.15	86.99±0.19	51.36±0.45	63.90±0.18	63.64±0.26	78.28±0.32
	MixOE [15]	89.78±0.15	89.15±0.22	43.64±1.04	96.63±0.04	96.15±0.06	14.07±0.25	70.31±0.32	65.02±0.37	74.29±0.40	66.31±0.16	65.63±0.21	85.06±0.26
	DivOE [13]	98.95±0.03	98.78±0.04	5.19±0.21	98.47±0.04	98.43±0.05	8.60±0.22	81.40±0.16	79.67±0.17	57.54±0.71	66.47±0.11	68.14±0.12	83.32±0.34
	MaCS	99.24±0.02	99.11±0.04	3.28±0.11	98.46±0.04	98.24±0.06	7.62±0.21	86.58±0.13	85.73±0.22	49.90±0.3	67.73±0.24	68.42±0.26	81.24±0.23
	MaCS*	**99.24±0.02**	**99.11±0.04**	**3.28±0.11**	**98.62±0.04**	**98.46±0.05**	**6.86±0.18**	**89.75±0.14**	**88.39±0.25**	**39.75±0.58**	**74.33±0.21**	**74.38±0.21**	**72.92±0.47**
LSUN-C	OE [3]	**99.74±0.01**	**99.74±0.01**	**1.10±0.07**	99.65±0.01	99.65±0.01	1.66±0.09	**97.00±0.08**	**96.94±0.07**	**14.94±0.61**	96.22±0.06	96.30±0.08	20.49±0.34
	Energy [30]	99.55±0.02	99.34±0.04	1.46±0.09	99.43±0.02	99.41±0.03	3.20±0.17	94.67±0.08	95.06±0.09	31.47±0.27	93.80±0.09	94.36±0.08	35.07±0.72
	MixOE [15]	97.30±0.09	97.07±0.12	11.92±0.25	97.99±0.03	97.65±0.05	9.36±0.15	92.08±0.14	91.27±0.16	31.39±0.56	91.67±0.05	90.61±0.09	30.07±0.29
	DivOE [13]	99.64±0.02	99.64±0.02	1.80±0.14	99.56±0.02	99.56±0.02	2.38±0.16	96.54±0.05	96.51±0.05	17.35±0.36	95.35±0.08	95.61±0.07	25.86±0.58
	MaCS	99.64±0.01	99.63±0.02	1.62±0.08	99.73±0.01	99.72±0.01	**1.11±0.06**	95.84±0.11	95.59±0.14	19.70±0.64	96.36±0.09	96.36±0.11	18.37±0.47
	MaCS*	99.64±0.01	99.63±0.02	1.62±0.08	**99.75±0.01**	**99.75±0.02**	1.16±0.09	96.03±0.10	95.72±0.14	18.39±0.71	**96.74±0.05**	**96.82±0.03**	**17.02±0.82**
SVHN	OE [3]	**99.35±0.03**	99.19±0.06	**2.14±0.08**	98.99±0.03	98.92±0.04	4.98±0.23	88.14±0.20	85.73±0.23	42.01±0.39	85.78±0.17	81.06±0.29	41.88±0.34
	Energy [30]	99.00±0.04	98.42±0.07	2.62±0.09	96.36±0.07	95.43±0.11	11.94±0.26	89.39±0.09	89.14±0.12	43.40±0.37	80.65±0.14	76.10±0.25	49.97±0.41
	MixOE [15]	91.68±0.19	90.71±0.27	31.18±0.97	89.52±0.09	84.02±0.18	28.69±0.30	74.84±0.28	67.66±0.34	63.05±0.44	76.51±0.23	68.05±0.30	53.54±0.62
	DivOE [13]	99.11±0.03	98.85±0.06	3.23±0.14	98.00±0.05	97.71±0.07	9.27±0.25	86.89±0.14	84.68±0.13	44.44±0.42	83.47±0.22	78.62±0.32	45.34±0.3
	MaCS	99.31±0.02	99.14±0.04	2.72±0.14	99.49±0.02	99.44±0.02	2.26±0.05	90.01±0.1	88.97±0.16	40.09±0.43	88.11±0.12	84.47±0.29	39.61±0.4
	MaCS*	99.31±0.02	**99.14±0.04**	2.72±0.14	**99.65±0.02**	**99.64±0.02**	**1.63±0.11**	**93.03±0.13**	**92.40±0.20**	**32.68±0.50**	**88.93±0.11**	**86.11±0.13**	**39.76±0.74**
Places365	OE [3]	96.73±0.07	96.86±0.08	14.74±0.4	95.03±0.08	95.00±0.09	21.76±0.54	86.42±0.25	85.56±0.26	50.97±0.9	83.72±0.16	82.56±0.23	55.82±0.45
	Energy [30]	**97.06±0.08**	**97.29±0.07**	15.50±0.52	94.39±0.06	94.46±0.07	24.72±0.32	82.93±0.08	82.61±0.10	58.95±0.49	79.80±0.18	78.96±0.20	61.28±0.52
	MixOE [15]	87.56±0.20	87.63±0.27	60.35±0.90	90.41±0.11	88.86±0.15	35.43±0.46	76.75±0.30	73.61±0.41	69.26±0.69	79.99±0.18	76.46±0.26	60.52±0.57
	DivOE [13]	96.34±0.11	96.41±0.10	16.16±0.54	94.82±0.04	94.80±0.05	22.20±0.37	85.18±0.10	84.44±0.07	53.36±0.42	83.16±0.17	82.13±0.17	56.95±0.52
	MaCS	97.03±0.08	97.24±0.07	13.00±0.45	95.98±0.07	96.08±0.08	18.96±0.34	86.56±0.14	85.67±0.18	51.04±0.74	84.26±0.2	82.96±0.21	54.02±0.59
	MaCS*	97.03±0.08	97.24±0.07	**13.00±0.45**	**96.43±0.12**	**96.59±0.11**	**16.89±0.61**	**87.23±0.20**	**86.41±0.29**	**50.86±0.75**	**84.88±0.20**	**83.87±0.19**	**53.65±0.58**

(a) CIFAR-10 and CIFAR-100

		SVHN						Imagenet32					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
Textures	OE [3]	99.75±0.00	99.79±0.00	0.43±0.03	99.96±0.00	99.96±0.00	0.02±0.01	80.16±0.03	7.24±0.03	75.17±0.09	84.27±0.04	82.08±0.04	71.51±0.12
	Energy [30]	99.71±0.00	99.74±0.00	0.54±0.03	99.97±0.00	99.97±0.00	0.01±0.0	86.78±0.02	85.81±0.03	67.39±0.10	76.57±0.05	74.65±0.04	82.13±0.08
	MixOE [15]	94.98±0.05	94.39±0.10	22.32±0.10	93.75±0.05	92.32±0.08	22.01±0.20	55.90±0.11	37.25±0.08	85.75±0.10	59.11±0.07	42.78±0.07	85.43±0.11
	DivOE [13]	99.85±0.00	99.86±0.01	0.10±0.01	99.98±0.0	99.98±0.0	0.01±0.01	86.59±0.03	85.18±0.04	67.12±0.11	85.41±0.03	84.01±0.04	70.45±0.08
	MaCS	99.87±0.0	99.88±0.0	0.11±0.01	99.95±0.00	99.96±0.00	0.00±0.00	87.43±0.04	86.19±0.04	64.63±0.20	83.56±0.03	81.44±0.04	73.58±0.13
	MaCS*	**99.91±0.00**	**99.92±0.00**	**0.00±0.00**	**99.99±0.00**	**99.99±0.00**	**0.00±0.00**	**87.43±0.04**	**86.19±0.04**	**64.63±0.20**	**88.73±0.03**	**87.46±0.06**	**63.80±0.17**
iSUN	OE [3]	100.00±0.00	100.00±0.	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.98±0.06	64.30±0.06	**73.99±0.09**	57.70±0.1	51.63±0.08	85.69±0.11
	Energy [30]	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.47±0.06	63.90±0.10	73.16±0.13	62.10±0.09	57.62±0.07	84.87±0.08
	MixOE [15]	98.31±0.04	97.86±0.07	6.89±0.13	97.15±0.03	96.31±0.05	10.85±0.1	64.99±0.11	54.92±0.10	71.72±0.16	57.13±0.08	49.72±0.06	86.42±0.11
	DivOE [13]	100.00±0.00	99.99±0.00	0.01±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.86±0.07	64.39±0.11	73.48±0.13	57.49±0.07	51.88±0.07	86.12±0.09
	MaCS	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	72.65±0.09	68.07±0.12	73.99±0.15	56.95±0.08	51.74±0.07	85.42±0.10
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**72.65±0.09**	**68.07±0.12**	73.99±0.15	**64.22±0.07**	**58.91±0.12**	**82.96±0.13**
LSUN-C	OE [3]	99.98±0.0	99.98±0.0	0.01±0.0	100.00±0.00	100.00±0.00	0.00±0.00	99.66±0.00	99.74±0.00	0.04±0.01	99.69±0.00	99.75±0.00	0.14±0.01
	Energy [30]	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	99.72±0.0	99.78±0.0	0.05±0.00	99.61±0.00	99.70±0.00	0.06±0.01
	MixOE [15]	98.31±0.04	97.86±0.07	6.89±0.13	96.38±0.03	95.72±0.04	14.25±0.14	88.62±0.06	86.01±0.10	38.45±0.12	93.30±0.03	93.07±0.04	30.82±0.08
	DivOE [13]	99.97±0.00	99.97±0.01	0.02±0.01	100.00±0.00	100.00±0.00	0.00±0.00	99.50±0.00	99.61±0.00	0.16±0.01	99.65±0.00	99.73±0.00	0.20±0.01
	MaCS	100.00±0.00	100.00±0.00	0.00±0.00	99.99±0.0	99.99±0.0	0.00±0.0	99.81±0.00	99.86±0.00	0.04±0.00	99.75±0.0	99.81±0.0	0.06±0.01
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**99.81±0.00**	**99.86±0.00**	**0.04±0.00**	**99.93±0.00**	**99.94±0.00**	**0.04±0.01**
Places365	OE [3]	99.99±0.00	9.99±0.00	0.01±0.0	100.00±0.00	100.00±0.00	0.00±0.00	93.00±0.06	94.77±0.04	49.38±0.37	96.47±0.03	97.18±0.02	23.34±0.28
	Energy [30]	99.98±0.00	99.98±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	97.43±0.03	98.15±0.03	15.88±0.45	97.01±0.03	97.78±0.02	20.76±0.34
	MixOE [15]	97.24±0.05	96.44±0.09	10.87±0.16	96.37±0.05	95.21±0.06	13.02±0.18	61.27±0.12	58.17±0.12	85.17±0.14	59.95±0.10	56.62±0.08	85.34±0.18
	DivOE [13]	99.99±0.00	99.99±0.00	0.02±0.01	100.00±0.00	100.00±0.00	0.00±0.00	95.26±0.02	96.50±0.02	36.82±0.19	97.32±0.02	97.88±0.02	16.51±0.12
	MaCS	99.99±0.0	99.99±0.0	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	97.57±0.02	98.21±0.02	14.65±0.31	97.75±0.02	98.24±0.02	13.34±0.21
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**97.57±0.02**	**98.21±0.02**	**14.65±0.31**	**99.36±0.02**	**99.46±0.02**	**0.92±0.05**

(b) SVHN and Imagenet-32

TABLE III: Optimum value of m reported while using different ID data and models.

ID Data \ Models	WRN	Allconv	Resnet-18	Densenet-121
CIFAR-10	0.5	0.9	0.8	0.7
CIFAR-100	0.8	0.9	0.8	0.8
SVHN	0.8	0.8	0.8	0.9
Imagenet-32	0.5	0.6	0.5	0.6

B. Training with Different Neural Networks

After achieving favorable outcomes of MaCS and MaCS* with WRN, and Allconv, we sought to determine if this performance could be replicated on other models. To this end, we trained each of the reference methods, as well as MaCS, under similar training configurations, but using different backbone architectures, namely Resnet-18 and Densenet-121, both of which are widely used image classification models. We utilized all four ID datasets and all five outlier datasets. We report the test results in Tables IVa, IVb and it is evident that our methods consistently achieved the best or second-best performance across the majority of the test datasets. These findings confirm that our method's performance is not limited to a particular type of neural network, as it demonstrates the capacity to achieve optimal results regardless of the network employed.

C. Detection Performance with and without Margin bound

In this ablation study, we eliminated the bounded margin and relied solely on MCD to check the influence of m to the overall performance. We trained a WRN backbone on all four ID datasets, and tested on all OOD datasets. The result is depicted in Fig. 5 as a bar chart, where we can observe a significant decline in the performance across all ID datasets

TABLE IV: Comprehensive OOD detection results obtained by training different ID datasets on different backbone architectures. Best, and second best results are represented in bold and underline respectively. For same results with other methods, we choose ours to be best or the second best.

Method	Models	CIFAR-10			CIFAR-100		
		AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↑
[3]	Allconv	97.90±0.03	97.86±0.04	10.18±0.19	83.07±0.12	81.83±0.18	52.11±0.18
OE	Resnet-18	97.51±0.04	**97.37±0.06**	11.96±0.16	86.39±0.15	84.56±0.19	47.35±0.40
	Densenet-121	96.71±0.06	96.29±0.10	14.86±0.25	83.95±0.19	81.42±0.31	52.68±0.38
[30]	Allconv	96.79±0.04	96.60±0.06	13.94±0.16	79.17±0.13	78.13±0.18	58.56±0.36
Energy	Resnet-18	97.46±0.06	97.31±0.1	12.74±0.27	85.27±0.12	85.35±0.15	55.37±0.26
	Densenet-121	96.89±0.05	96.70±0.08	14.59±0.26	82.16±0.12	81.81±0.20	63.99±0.33
[15]	Allconv	93.52±0.04	91.77±0.08	22.83±0.22	77.89±0.11	74.39±0.15	60.10±0.33
MixOE	Resnet-18	84.95±0.13	81.51±0.18	48.40±0.35	77.30±0.22	72.54±0.33	61.74±.38
	Densenet-121	85.10±0.12	83.87±.15	57.69±0.38	74.18±.1	71.43±0.15	72.35±0.25
[13]	Allconv	97.70±0.03	97.65±0.04	11.24±0.13	81.96±0.12	81.10±0.13	54.30±0.27
DivOE	Resnet-18	97.12±0.06	96.93±0.09	13.64±0.18	85.25±0.12	83.30±0.20	50.24±0.29
	Densenet-121	96.33±0.06	95.98±0.11	16.58±0.30	84.00±0.12	82.02±0.18	53.76±0.25
	Allconv	98.39±0.03	98.35±0.04	7.91±0.1	83.66±0.13	82.53±0.2	50.39±0.31
MaCS	Resnet-18	97.00±0.05	96.56±0.10	13.29±0.22	87.39±0.13	**86.27±0.16**	47.12±0.27
	Densenet-121	95.99±0.05	95.59±0.08	18.35±0.30	83.31±0.10	81.75±0.16	56.88±0.19
	Allconv	**98.63±0.04**	**98.63±0.04**	**6.85±0.18**	**85.74±0.12**	**84.90±0.09**	**48.11±0.51**
MaCS*	Resnet-18	**97.61±0.03**	97.28±0.05	**10.69±0.13**	**87.47±0.1**	85.43±0.13	**44.80±0.27**
	Densenet-121	**97.58±0.04**	**97.33±0.08**	**11.37±0.16**	**85.37±0.12**	**82.22±0.19**	**49.89±0.27**

(a) CIFAR-10 and CIFAR-100

Method	Models	SVHN			Imagenet-32		
		AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↑
[3]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.63±0.03	86.13±0.03	36.14±0.06
OE	Resnet-18	99.99±0.00	99.99±0.00	0.00±0.00	92.05±0.03	91.01±0.04	27.79±.04
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	91.19±0.02	90.37±0.02	32.39±0.09
[30]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.06±0.03	85.95±0.02	37.56±0.08
Energy	Resnet-18	99.97±0.00	99.97±0.00	0.01±0.00	90.23±0.03	89.30±0.03	31.11±0.06
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	92.77±0.02	92.27±0.02	30.89±0.08
[15]	Allconv	95.91±0.04	94.89±0.06	15.03±0.13	73.75±0.05	68.14±0.05	58.09±0.08
MixOE	Resnet-18	92.07±0.08	88.80±0.12	24.21±0.16	77.98±0.04	69.96±0.08	55.45±0.08
	Densenet-121	94.99±0.05	93.71±0.06	18.49±0.13	75.10±0.06	67.30±0.10	63.11±0.1
[13]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.97±0.02	86.70±0.02	34.66±0.03
DivOE	Resnet-18	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	92.61±0.02	91.65±0.03	**27.34±0.06**
	Densenet-121	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	91.35±0.02	90.59±0.03	31.47±0.08
	Allconv	99.99±0.00	99.99±0.00	0.00±0.00	87.60±0.02	86.25±0.02	34.48±0.06
MaCS	Resnet-18	99.99±0.00	99.99±0.00	0.00±0.00	92.81±0.02	92.29±0.02	29.13±0.06
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	93.26±0.03	92.47±0.03	27.00±0.08
	Allconv	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**90.45±0.02**	**89.15±0.04**	**29.55±0.06**
MaCS*	Resnet-18	99.99±0.00	99.99±0.00	0.00±0.00	**92.81±0.02**	**92.29±0.02**	29.13±0.06
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	**94.40±0.2**	**92.84±0.02**	**25.94±0.06**

(b) SVHN and Imagenet-32

when MaCS is not subjected to a margin bound. Although MCD assigns a penalty of zero to score explosions, it is evident that these values remain ambiguous and do not contribute to learning when not substituted with a specific weight, which is the value of m in this instance. In essence, when one considers (4), and when margin is not used, $\mathcal{W}_{MaCS}$ will either assume a value of 0 or simply the MCD value, which may be null or the difference between ID and OOD scores. However, this difference does not correspond to the desired difference that is obtainable with margin.

VIII. Conclusion

In this paper, we introduced a novel and straightforward methodology aimed at improving OOD detection by establishing a compact decision boundary between ID and OOD data. To this end, we recognized a disguised OOD detection problem that existed in OE setting, i.e., score explosions, and proposed a solution, MaCS which first penalizes score explosions, and then substitutes it with a margin value to realize the difference between ID and OOD data to be as large as possible. Our approach significantly enhanced the OOD detection and pro-

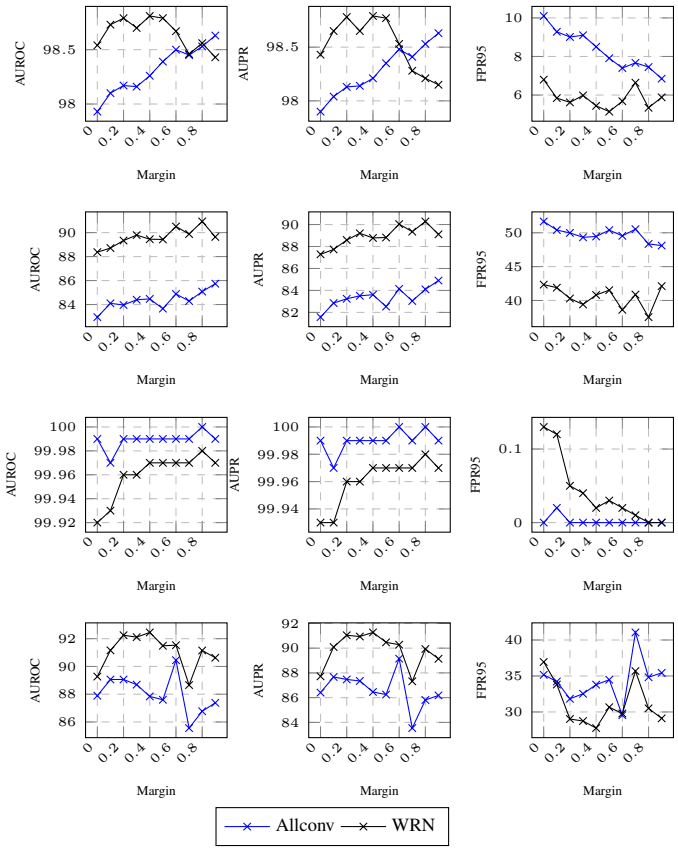

Fig. 4: Line graph representing the OOD detection performance of MaCS across different margin values. Each row represents different ID datasets in the order from top to bottom: CIFAR-10, CIFAR-100, SVHN, Imagenet-32. The results represent an average value over multiple OOD datasets.

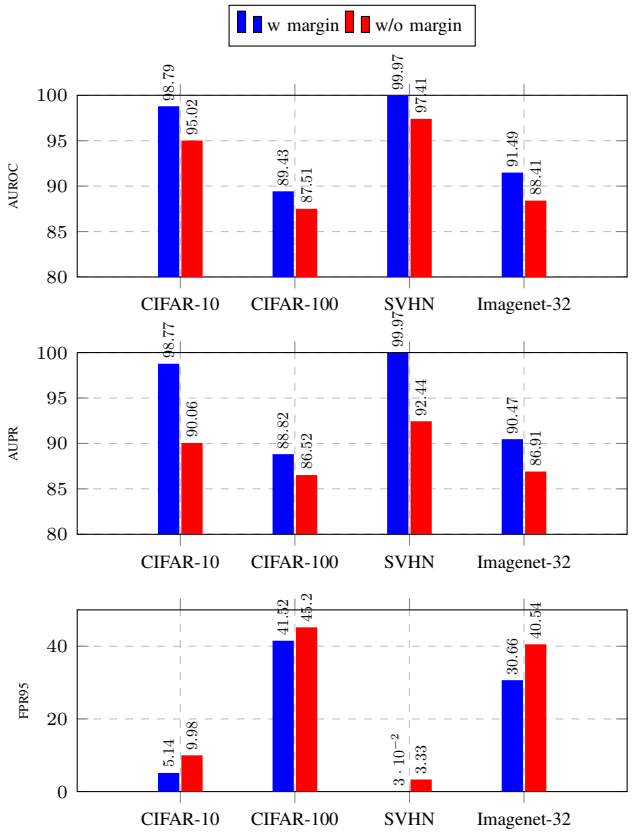

Fig. 5: Bar-graph representing different OOD metrics for individual ID datasets with and without margin bound. A WRN model was trained on these ID datasets.

vided competitive performance when compared with several S.O.T.A benchmarks across four ID datasets, and five OOD datasets in the image classification domain. Importantly, our proposed method was also able to achieve significant gain in ID accuracy. To summarize the detection performance, our method exhibited a remarkable gain in AUROC, AUPR, and FPR95, reaching a maximum improvement of 2.73%, 3.26%, and 9.06%, respectively. These results affirm its effectiveness and thus demonstrate the synergy of OE with our method in advancing the field of OOD detection.

Reproducibility: All of our source codes, pretrained models, and results are publicly shared at: https://github.com/lakpa-tamang9/margin_ood

Acknowledgement: We thank the anonymous reviewers for their valuable comments, which improved the paper. Mr Lakpa Tamang is supported by the Deakin University Postgraduate Research (DUPR) Scholarship. Dr Mohamed Reda Bouadjenek and Dr Sunil Aryal are supported by the Air Force Office of Scientific Research (AFOSR) under award number FA2386-23-1-4003.

REFERENCES

[1] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.

[2] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2016.

[3] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations*, 2018.

[4] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," 2015.

[5] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32598–32611, 2022.

[6] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[8] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631–636, 2003.

[9] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[10] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[11] N. Japkowicz, C. Myers, M. Gluck, *et al.*, "A novelty detection approach to classification," in *IJCAI*, vol. 1, pp. 518–523, Citeseer, 1995.

[12] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pp. 430–445, Springer, 2021.

[13] J. Zhu, Y. Geng, J. Yao, T. Liu, G. Niu, M. Sugiyama, and B. Han, "Diversified outlier exposure for out-of-distribution detection via informative extrapolation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] F. Zhu, Z. Cheng, X.-Y. Zhang, and C.-L. Liu, "Openmix: Exploring outlier samples for misclassification detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12074–12083, 2023.

[15] J. Zhang, N. Inkawhich, R. Linderman, Y. Chen, and H. Li, "Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5531–5540, 2023.

[16] Q. Wang, J. Ye, F. Liu, Q. Dai, M. Kalander, T. Liu, H. Jianye, and B. Han, "Out-of-distribution detection with implicit outlier transformation," in *The Eleventh International Conference on Learning Representations*, 2023.

[17] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *International Conference on Machine Learning*, pp. 8759–8773, PMLR, 2022.

[18] S. Jung, J. Lee, D. Gwak, S. Choi, and J. 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*, pp. 15425–15434, 2021.

[19] Z. Zhang and X. Xiang, "Decoupling maxlogit for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3388–3397, 2023.

[20] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018.

[21] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

[22] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[23] Y. Sun and Y. Li, "Dice: Leveraging sparsification for out-of-distribution detection," in *European Conference on Computer Vision*, pp. 691–708, Springer, 2022.

[24] F. Pinto, H. Yang, S. N. Lim, P. Torr, and P. Dokania, "Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14608–14622, 2022.

[25] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

[26] K. LEE, K. Lee, H. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *ICLR 2018*, ICLR 2018, 2018.

[27] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5216–5223, 2020.

[28] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," *Neurocomputing*, vol. 441, pp. 138–150, 2021.

[29] H. Zheng, Q. Wang, Z. Fang, X. Xia, F. Liu, T. Liu, and B. Han, "Out-of-distribution detection learning with unreliable out-of-distribution sources," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21464–21475, 2020.

[31] P. Morteza and Y. Li, "Provable guarantees for understanding out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 7831–7840, 2022.

[32] D. Choi and D. Na, "Towards reliable ai model deployments: Multiple input mixup for out-of-distribution detection," *arXiv preprint arXiv:2312.15514*, 2023.

[33] S. Yue and T. Wang, "Imbalanced malware images classification: a cnn based approach," *arXiv preprint arXiv:1708.08042*, 2017.

[34] K. R. M. Fernando and C. P. Tsokos, "Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2940–2951, 2021.

[35] A. Anand, G. Pugalenthi, G. B. Fogel, and P. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino acids*, vol. 39, pp. 1385–1391, 2010.

[36] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[37] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[38] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, p. 7, Granada, Spain, 2011.

[39] P. Chrabaszcz, I. Loshchilov, and F. Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[41] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[42] V. U. Prabhu and A. Birhane, "Large image datasets: A pyrrhic win for computer vision?," *arXiv preprint arXiv:2006.16923*, 2020.

[43] 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.

[44] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[45] B. Zhou, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *Journal of Vision*, vol. 17, no. 10, pp. 296–296, 2017.

[46] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference 2016*, British Machine Vision Association, 2016.

[47] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.